GAPSEER: HYPERBOLIC SEMANTIC AXIS RESPONSES FOR WEAKLY SUPERVISED FINE-GRAINED VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing weakly supervised fine-grained video anomaly detection (WS-FVAD) methods typically construct category responses based on cross-modal semantic similarity between concrete event instances and category texts describing abstract anomaly semantics. However, the semantic granularity mismatch between abstract category descriptions and diverse visual realizations makes similarity-based responses unreliable. Although hyperbolic entailment cones provide an asymmetric mechanism for modeling instance-to-category relations, visually similar patterns shared by different anomaly categories may cause an event instance to fall into multiple category cones, resulting in ambiguous category responses. To address this issue, we propose GapSeer, which constructs fine-grained category responses by modeling how strongly a video segment specializes along each category-specific semantic direction, rather than relying on overall visual-text similarity or cone inclusion alone. Specifically, the Hyperbolic Semantic Axis (HSA) defines a directed semantic trajectory from abstract category semantics toward concrete event realizations, while the Signed Axial Response (SAR) quantifies the signed progression of each video segment along the corresponding axis. This formulation accommodates diverse visual manifestations within the same anomaly category while distinguishing visually similar but semantically different events, thereby reducing ambiguity among competing category responses. Extensive experiments on the XD-Violence and UCF-Crime benchmarks demonstrate the effectiveness of our method.

1 INTRODUCTION

Weakly Supervised Video anomaly Detection (WS-VAD) aims to temporally localize anomalous events in untrimmed videos using only video-level labels (Sultani et al., 2018; Tian et al., 2021; Chen et al., 2023). Building on recent advances in vision-language models (Radford et al., 2021; Zhou et al., 2022b;a), research has increasingly moved toward weakly supervised fine-grained video anomaly detection (WS-FVAD) (Wu et al., 2024; Zhang et al., 2026; Huang et al., 2025). Whereas conventional WS-VAD typically distinguishes anomalous from normal video segments, WS-FVAD additionally identifies the corresponding anomaly category. Such category-level information is particularly valuable for public-safety and traffic-monitoring applications, where different anomaly types call for different response measures. Existing WS-FVAD methods (Wu et al., 2024; Huang et al., 2025; Zhang et al., 2026) typically leverage pretrained vision-language models by dividing a video into temporal segments and measuring the cross-modal cosine similarity between each segment and textual representations of anomaly categories. The resulting similarity scores are then treated as segment-level category responses.

When computing cross-modal cosine similarity, the textual modality represents abstract category-level semantics, whereas the visual modality encodes concrete event instances (Ramasinghe et al., 2024). This discrepancy introduces a semantic granularity mismatch between the two modalities, as shown in Fig. 1(a). However, as shown in Fig. 1(b) cosine similarity primarily measures the proximity between their semantic representations, but it is insufficient to capture the instance-to-category membership relation required by our task. Specifically, a video clip depicts a particular event instance and typically contains rich contextual details, including the surrounding scene, temporal ac-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

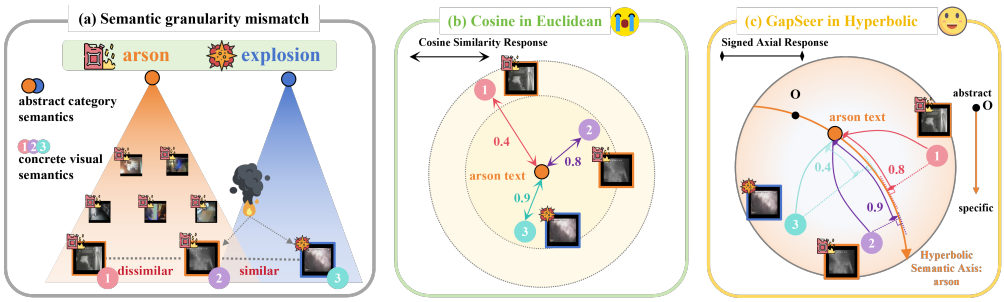

Figure 1: **Motivation for GapSeer.** (a) Abstract category texts and concrete video events differ in semantic granularity. (b) Euclidean cosine similarity can produce weak correct-category responses and strong responses to visually similar events from other categories. (c) GapSeer constructs category responses from signed progress along category-specific hyperbolic semantic axes.

tion dynamics, participating entities, and their interactions, while the corresponding category text only summarizes the core semantics shared by a class of abnormal events. Therefore, a mismatch in similarity between a video segment and its category text may simply arise from instance-specific details rather than indicating a category mismatch.

These observations reveal a relation that similarity alone cannot capture, where a concrete event instance may instantiate an abstract category without being highly similar to its textual representation. Hyperbolic entailment cones provide a natural way to model such asymmetric instance-to-category relations, allowing an abstract concept to subsume visually diverse realizations (Ganea et al., 2018; Desai et al., 2023; Pal et al., 2025). However, visually shared content across abnormal categories may cause the same instance to fall within multiple category cones. In such cases, cone inclusion indicates only semantic compatibility and provides limited discrimination among competing categories. This limitation is particularly critical in WS-FVAD, where instances of the same abnormal category may exhibit highly diverse visual realizations, while different abnormal categories can share similar visual content. An effective category response should therefore accommodate instance-specific variations within a category while remaining discriminative among visually similar but semantically distinct events. This requires going beyond whether an instance lies within a category cone and further characterizing how strongly it progresses along each category-specific semantic direction. Motivated by this observation, we propose GapSeer, which constructs category responses by measuring how a video segment develops along each category-specific semantic direction, rather than relying on its overall similarity to the category text (Fig. 1(c)). Specifically, we introduce the Hyperbolic Semantic Axis (HSA), which represents each category as a directed semantic trajectory rooted at its abstract textual concept. This formulation allows visually diverse instances to be characterized according to how they realize the same underlying category semantics. Building on HSA, the Signed Axial Response (SAR) quantifies the signed progression of a video segment along the corresponding semantic direction. As a result, visually different segments can still yield consistent responses when they instantiate the same abnormal category, while visually similar but semantically distinct events can be differentiated according to their progression along different category-specific axes. Evaluating each segment against all semantic axes produces a snippet-level alignment map that captures category-specific semantic development rather than overall visual-text similarity.

The main contributions of this work are summarized as follows: 1) We reveal the semantic granularity mismatch between concrete video instances and abstract category texts, showing that existing cosine similarity is insufficient to model the asymmetric instance-to-category relation. 2) We propose GapSeer, which models category-specific semantic progression through HSA and SAR, accommodating intra-class diversity while reducing confusion among visually similar categories. 3) We conduct a comprehensive benchmark analysis of our proposed approach on the widely used XD-Violence and UCF-Crime datasets. The experimental results indicate that our introduced Hyperbolic semantic axis response brings the current state-of-the-art (SOTA) best suited for WS-FVAD tasks.

2 Related Work

2.1 Weakly Supervised Fine-Grained Video Anomaly Detection

WS-VAD focuses on distinguishing anomalous snippets from normal ones under video-level supervision (Zhong et al., 2019; Feng et al., 2021; Lv et al., 2023; Zhang et al., 2023; Chen et al., 2024; Yang et al., 2024), whereas WS-FVAD further identifies the categories of anomalous events (Wu et al., 2023; 2024; Zhang et al., 2026). Recent WS-FVAD methods commonly infer anomaly categories by aligning video snippets with category texts under video-level supervision. VadCLIP (Wu et al., 2024) constructs a snippet-category alignment map from visual-text similarity and learns category predictions through MIL. Subsequent methods improve the semantic evidence used in this alignment. LEC-VAD (Wang & Chen, 2025) enriches concise category descriptions with memory-bank prototypes, while Ex-VAD (Huang et al., 2025) combines generated event explanations with augmented category labels. On the visual and learning side, DSANet (Yin et al., 2026) separates event-centric from background-centric features to reduce category confusion, and Fine-VAD (Zhang et al., 2026) progressively guides representation learning from generic anomaly patterns to category-specific semantics. These methods strengthen the representations and supervision involved in alignment, but the resulting category responses are still based on overall visual-text similarity. Consequently, visual details that distinguish a concrete event from its abstract category description can also influence its category response.

2.2 Hyperbolic Entailment Cones

Hyperbolic embeddings provide a geometric framework for representing hierarchical structure (Nickel & Kiela, 2017; 2018). The relation between general concepts and specific instances has also been studied through hyperbolic entailment cones. Ganea et al. (2018) use nested geodesically convex cones to represent hierarchical partial orders, and Shadow Cones (Yu et al., 2024) extends this idea to a broader family of cone constructions. In video understanding, Long et al. (2020) learn hyperbolic action prototypes and use entailment cones to retrieve hierarchically related actions. In vision-language learning, MERU (Desai et al., 2023) combines Lorentz-distance-based image-text matching with a cone-based loss that encourages text-to-image entailment. HyCoCLIP (Pal et al., 2025) further applies entailment supervision to compositional relations among images, image regions, and their textual descriptions. PHyCLIP (Yoshikawa & Matsubara, 2026) models hierarchy and compositionality through an ℓ_1-product of hyperbolic factors, while Alignment across Trees (Wu et al., 2026) constructs and aligns visual and textual feature hierarchies on hyperbolic manifolds with different curvatures. In MERU and HyCoCLIP, cone constraints organize the hierarchy of learned representations alongside contrastive image-text alignment. However, cone inclusion only indicates semantic compatibility between instances and categories, which is insufficient to distinguish visually similar anomaly categories. This motivates modeling the degree to which an instance specializes along category-specific semantic directions.

3 Method

3.1 Overview

Problem formulation. WS-FVAD aims to localize anomalous events and identify their categories using only video-level supervision. Given an untrimmed video v, we denote its binary anomaly label by $y^b \in \{0,1\}$ and its multi-hot category label by $\mathbf{y}^m \in \{0,1\}^C$, where C is the total number of categories, with $c = 0$ denoting the normal category and $c \in \{1, \dots, C-1\}$ denoting the anomaly categories. A normal video has $y^b = 0$ and $y^m_0 = 1$; an anomalous video has $y^b = 1$, with $y^m_c = 1$ for each anomaly category present in v.

Overall framework. As illustrated in Fig. 2, the video encoder E_V extracts features $\mathbf{F}_v = \{\mathbf{f}_t\}_{t=0}^{T-1}$ from the T temporal snippets of v. The frozen CLIP text encoder E_T produces category features $\mathbf{F}_{\text{text}} = \{\mathbf{e}_c\}_{c=0}^{C-1}$. We map both modalities to the d-dimensional Lorentz manifold $\mathcal{L}^d_k$ with curvature $-1/k$, where $k > 0$. Following MERU (Desai et al., 2023), modality-specific scaling factors α_{video} and α_{text} control feature magnitudes before exponential mapping:

$$\mathbf{H} = \{\mathbf{h}_t\}_{t=0}^{T-1}, \quad \mathbf{h}_t = \exp^k_{\boldsymbol{o}}([0, \alpha_{\text{video}}\mathbf{f}_t]); \qquad \mathbf{T} = \{\mathbf{t}_c\}_{c=0}^{C-1}, \quad \mathbf{t}_c = \exp^k_{\boldsymbol{o}}([0, \alpha_{\text{text}}\mathbf{e}_c]), \quad (1)$$

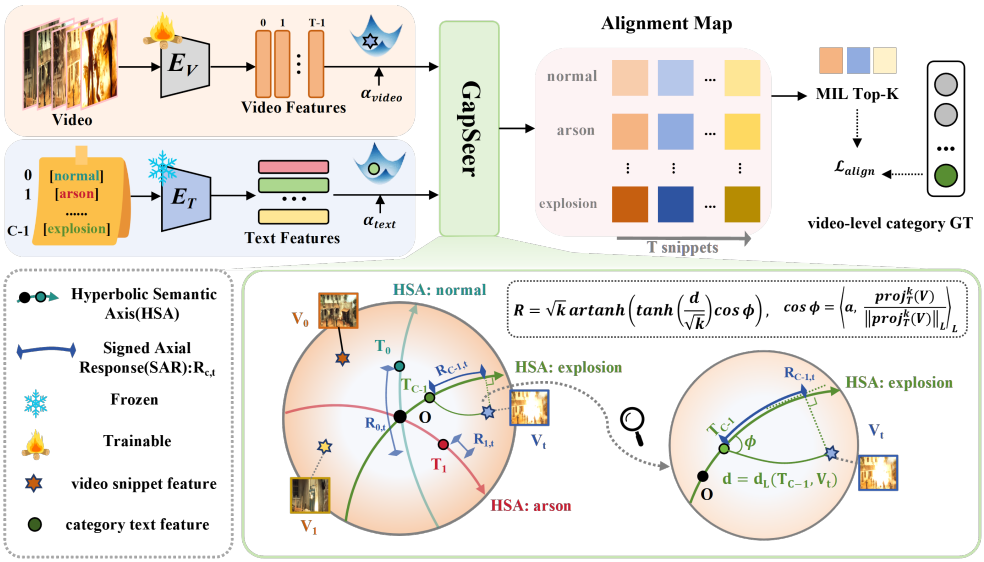

Figure 2: **Overview of GapSeer.** HSA defines an oriented geodesic through the Lorentz origin and each category text representation, taking the direction away from the origin as positive to specify how abstract category semantics develop toward concrete visual instances. SAR uses the Lorentz distance between a video snippet and the category text, together with their angle at the text representation, to measure signed progress along the axis. The SARs computed for each video snippet on every HSA form the alignment map.

where $\boldsymbol{o} = [\sqrt{k}, \mathbf{0}]$ is the Lorentz origin, $\mathbf{0} \in \mathbb{R}^d$, and $\exp_{\boldsymbol{o}}^k$ maps tangent vectors at $\boldsymbol{o}$ onto $\mathcal{L}_k^d$. Appendix A provides the geometric definitions and formulas.

GapSeer uses HSA and SAR to construct the visual-text alignment map $\mathbf{M}$. MIL top-K pooling aggregates its responses for video-level supervision.GapSeer can thus be incorporated into WS-FVAD frameworks using visual-text alignment.

3.2 Hyperbolic Semantic Axis

As discussed in Sec. 1, a useful category response in WS-FVAD should accommodate intra-class visual diversity while discriminating visually similar categories. Neither proximity to a category text nor semantic compatibility alone explicitly captures how a snippet instantiates one category rather than another. Motivated by the abstraction-to-instance organization of hyperbolic vision-language representations (Desai et al., 2023; Pal et al., 2025), we use each category text as a semantic anchor and orient its radial geodesic away from the origin to define a category-specific direction toward concrete visual realizations.

For category c, let $\mathbf{t}_c \in \mathcal{L}_k^d$ denote its hyperbolic text representation. We define the outward unit tangent $\mathbf{a}_c \in T_{\mathbf{t}_c}\mathcal{L}_k^d$ as

$$\mathbf{a}_c = -\frac{\operatorname{proj}_{\mathbf{t}_c}^k(\boldsymbol{o})}{\left\|\operatorname{proj}_{\mathbf{t}_c}^k(\boldsymbol{o})\right\|_{\mathcal{L}}}, \tag{2}$$

where the minus sign reverses the direction toward the origin, yielding the outward orientation. Starting from $\mathbf{t}_c$ along $\mathbf{a}_c$, the corresponding geodesic is parameterized by

$$\gamma_c(u) = \exp_{\mathbf{t}_c}^k(u\mathbf{a}_c), \qquad u \in \mathbb{R}. \tag{3}$$

We define this oriented geodesic as the **Hyperbolic Semantic Axis (HSA)** of category c, with $u > 0$ pointing from the abstract category concept toward concrete realizations. HSA specifies a category-specific semantic direction without requiring visual instances to lie on the axis or approach the category text. SAR quantifies signed progression along this axis (Sec. 3.3).

3.3 Signed Axial Response

Given an HSA, we next quantify how far a video snippet progresses along the corresponding category-specific direction. Distance alone cannot distinguish whether a snippet develops along or against the axis, while direction alone does not capture the extent of its progression. We therefore define the **Signed Axial Response (SAR)** as the signed axial displacement of a video snippet with respect to the category text, jointly determined by their Lorentz distance and relative angle through the hyperbolic right-triangle relation.

Derivation. For a video snippet representation $\mathbf{h}_t$ and a category text representation $\mathbf{t}_c$, let $d_{c,t} = d_{\mathcal{L}}(\mathbf{t}_c, \mathbf{h}_t)$ denote their Lorentz geodesic distance. For $d_{c,t} > 0$, the angle $\phi_{c,t}$ at $\mathbf{t}_c$ between the outward axis direction $\mathbf{a}_c$ and the geodesic direction toward $\mathbf{h}_t$ satisfies

$$\cos\phi_{c,t} = \left\langle \mathbf{a}_c, \frac{\mathrm{proj}^k_{\mathbf{t}_c}(\mathbf{h}_t)}{\left\|\mathrm{proj}^k_{\mathbf{t}_c}(\mathbf{h}_t)\right\|_{\mathcal{L}}} \right\rangle_{\mathcal{L}}. \tag{4}$$

Let $R_{c,t}$ denote the signed displacement from $\mathbf{t}_c$ to the orthogonal projection of $\mathbf{h}_t$ onto the oriented axis γ_c, with its sign determined by the orientation of γ_c. The geodesic segment from $\mathbf{t}_c$ to $\mathbf{h}_t$, its axial projection, and the orthogonal deviation form a hyperbolic right triangle. Applying the hyperbolic law of cosines to this right triangle yields

$$\tanh\left(\frac{R_{c,t}}{\sqrt{k}}\right) = \tanh\left(\frac{d_{c,t}}{\sqrt{k}}\right)\cos\phi_{c,t}. \tag{5}$$

Therefore,

$$R_{c,t} = \begin{cases} 0, & d_{c,t} = 0, \\ \sqrt{k}\,\mathrm{artanh}\left[\tanh\left(\dfrac{d_{c,t}}{\sqrt{k}}\right)\cos\phi_{c,t}\right], & d_{c,t} > 0. \end{cases} \tag{6}$$

Positive responses indicate that the visual semantics develop along the category's specialization direction, whereas negative responses indicate development along the opposite direction. Finally, the SARs computed for each video snippet on every HSA form the visual-text alignment map

$$\mathbf{M} = \left[\frac{R_{c,t}}{\tau}\right] \in \mathbb{R}^{C\times T}, \tag{7}$$

where $\tau > 0$ is the temperature used to scale the responses.

3.4 Training and Inference

Training. We adopt Fine-VAD (Zhang et al., 2026) without its intermediate level and use GapSeer alignment maps at both coarse and fine stages. Coarse supervision applies BCE to top-K-pooled snippet anomaly probabilities; fine supervision applies category cross-entropy after top-K pooling the SAR logits and softmax normalization. Both objectives backpropagate through SAR:

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{bce}} + \mathcal{L}_{\mathrm{ce}}. \tag{8}$$

Inference. We retain anomaly categories with video-level activations above θ_v and merge consecutive snippets whose softmax category probabilities exceed θ_s. Candidate instances are scored using the outer-inner criterion (Shou et al., 2018) and filtered by non-maximum suppression.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate GapSeer on two widely used WS-FVAD benchmarks: XD-Violence (Wu et al., 2020) and UCF-Crime (Sultani et al., 2018). XD-Violence is a large-scale surveillance video dataset containing diverse real-world anomalous events with both visual and audio information, while UCF-Crime consists of long untrimmed surveillance videos covering multiple types of criminal activities.

Following prior WS-FVAD work (Wu et al., 2024), we report mean Average Precision (mAP) at temporal Intersection-over-Union (IoU) thresholds from 0.1 to 0.5 in increments of 0.1, together with the average across these five thresholds (AVG).

Table 1: Comparison with existing WS-FVAD methods and evaluation of GapSeer across frameworks. Bold and underlined values indicate the best and second-best results, respectively.

Method	XD-Violence (mAP@IoU%)						UCF-Crime (mAP@IoU%)					
	0.1	0.2	0.3	0.4	0.5	AVG	0.1	0.2	0.3	0.4	0.5	AVG
AVVD TMM'23	30.51	25.75	20.18	14.83	9.79	20.21	10.27	7.01	6.25	3.42	3.29	6.05
ITC TIP'24	40.83	32.80	25.42	19.65	15.47	26.83	13.54	9.24	7.45	5.46	3.79	7.90
ReFLIP TCSVT'24	39.24	33.45	27.71	20.86	17.22	27.36	14.23	10.34	9.32	7.54	6.81	9.62
Ex-VAD ICML'25	40.14	32.75	28.78	20.15	18.35	28.23	16.51	12.35	9.41	7.82	4.65	10.15
Fine-VAD CVPR'26	44.06	<u>36.69</u>	<u>31.22</u>	25.80	21.58	31.87	21.43	**18.62**	**16.25**	<u>11.41</u>	7.26	<u>14.99</u>
VadCLIP AAAI'24	37.03	30.84	23.38	17.90	14.31	24.70	11.72	7.83	6.40	4.53	2.93	6.68
VadCLIP + **GapSeer**	38.47	32.13	26.02	20.13	15.74	26.50 (+1.80)	16.21	11.81	8.63	6.83	4.56	9.61 (+2.93)
DSANet AAAI'26	40.93	34.63	28.21	22.70	17.89	28.87	21.39	14.96	11.74	8.98	8.00	13.01
DSANet + **GapSeer**	42.82	36.54	31.01	25.35	19.81	31.11 (+2.24)	<u>22.75</u>	<u>17.92</u>	13.53	10.32	9.62	14.83 (+1.82)
Baseline	<u>45.51</u>	36.19	30.96	<u>26.23</u>	<u>22.50</u>	<u>32.28</u>	15.41	12.28	11.43	10.79	<u>9.64</u>	11.91
Baseline + GapSeer	**45.82**	**38.90**	**33.66**	**28.01**	**23.76**	**34.03** (+1.75)	**23.80**	16.85	<u>14.92</u>	**13.64**	**12.80**	**16.40** (+4.49)

4.2 Implementation Details

We follow VadCLIP (Wu et al., 2024) for the video encoder and use the frozen CLIP text encoder for category prompts. All models are trained for 10 epochs using AdamW on a single NVIDIA RTX A6000 GPU, with a batch size of 64 and a learning rate of 1×10^{-5}. We set $k = 1$. The initialization multipliers for the learnable visual and textual scaling factors are $(1.6, 0.3)$ on UCF-Crime and $(2.5, 0.5)$ on XD-Violence. We set $\tau = 0.02$ for both datasets and inference thresholds $\theta_v = 0.2$ and $\theta_s = 0.35$.

4.3 Comparison with State-of-the-Art Methods

We evaluate GapSeer's fine-grained detection performance and its applicability across WS-FVAD frameworks on XD-Violence and UCF-Crime.

Main results. We compare GapSeer with representative WS-FVAD methods (Wu et al., 2023; Liu et al., 2024; Dev et al., 2024; Wu et al., 2024; Huang et al., 2025; Yin et al., 2026; Zhang et al., 2026) on XD-Violence and UCF-Crime. As shown in Table 1, GapSeer raises the baseline's AVG mAP from 32.28% to 34.03% on XD-Violence and from 11.91% to 16.40% on UCF-Crime, achieving the highest AVG mAP on both benchmarks. On XD-Violence, GapSeer achieves the best mAP at every IoU threshold, with its advantage extending to the stricter thresholds of 0.4 and 0.5. On UCF-Crime, it reaches 12.80% mAP at IoU 0.5, compared with 7.26% for Fine-VAD. These results show that category-axis responses improve fine-grained detection and yield stronger localization under strict temporal overlap criteria.

Cross-framework applicability. We further apply GapSeer to VadCLIP and DSANet to assess its applicability across WS-FVAD frameworks using visual-text alignment. As shown in Table 1, GapSeer improves the AVG mAP of both models on both datasets. On XD-Violence, the gains are 1.80 for VadCLIP and 2.24 for DSANet; on UCF-Crime, they are 2.93 and 1.82, respectively. These results demonstrate that GapSeer's hyperbolic semantic axis responses generalize across different WS-FVAD architectures.

4.4 Ablation Studies

We conduct two ablation studies on UCF-Crime to examine the contributions of hyperbolic geometry and the SAR response function.

Effect of the Response Function. To assess how modeling instance-to-category relations affects performance, Table 2 compares SAR with five alternatives under the same Lorentz representation. The entailment-cone response achieves 6.51% AVG mAP, indicating that compatibility alone provides limited category discrimination. Radial-distance difference achieves 10.41%, suggesting that abstract-to-concrete progression is useful, but radial change does not capture category-specific di-

Table 2: Ablation study of response functions on UCF-Crime. Angular and axis-distance responses measure direction and proximity relative to the semantic axis, respectively; Lorentz-distance, entailment-cone, and radial-distance-difference responses evaluate text-snippet distance, cone compatibility, and radial displacement.

Response function	**mAP@IoU (%)**					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**AVG**
Angular response	6.55	6.54	6.54	6.51	6.39	6.51
Axis-distance response	6.54	6.53	6.49	6.45	6.29	6.46
Lorentz-distance response	6.75	6.75	6.75	6.71	6.67	6.73
Entailment-cone response	6.55	6.55	6.54	6.51	6.39	6.51
Radial-distance-difference response	11.81	11.43	10.56	9.14	9.12	10.41
SAR	**23.80**	**16.85**	**14.92**	**13.64**	**12.80**	**16.40**

Table 3: Ablation study of embedding geometry on UCF-Crime. The Euclidean axial response implements our proposed alignment principle in Euclidean space.

Alignment formulation	**mAP@IoU (%)**					
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**AVG**
Baseline	15.41	12.28	11.43	10.79	9.64	11.91
Euclidean axial response	19.06	15.25	14.20	12.85	11.72	14.62
SAR	**23.80**	**16.85**	**14.92**	**13.64**	**12.80**	**16.40**

rections. SAR reaches 16.40%, outperforming all alternatives at every IoU threshold. These results support the central design of GapSeer: category responses should capture not only concept-instance compatibility, but also the extent of progression along each category's semantic direction.

Effectiveness of Hyperbolic Geometry. To verify whether the improvement comes from hyperbolic geometry rather than simply replacing cosine similarity with a directed response, Table 3 compares cosine similarity, the Euclidean axial response, and SAR. The Euclidean axial response improves the AVG mAP from 11.91% to 14.62%, showing that measuring the displacement of a concrete event along a category-specific semantic direction provides a more effective alignment signal than overall visual-text similarity. Placing the same response principle in Lorentz space further increases the AVG mAP to 16.40% and improves the result at every IoU threshold. These results demonstrate that the gain comes from combining semantic specialization along category-specific directions with hyperbolic geometry.

4.5 Qualitative Results

In this section, we analyze GapSeer's ability to accommodate intra-class visual diversity and distinguish visually similar anomaly categories.

Fine-Grained Curves. To examine the effect of GapSeer on temporal category responses, Fig. 3 compares the prediction curves of the baseline, VadCLIP, and DSANet with their GapSeer variants. GapSeer produces clearer category responses around the anomalous intervals, with the most evident improvement appearing in the VadCLIP example. Similar refinements are also observed for the baseline and DSANet, showing that GapSeer improves fine-grained prediction curves across different WS-FVAD frameworks.

Category Responses under Low Visual-Text Similarity. To examine whether GapSeer preserves the correct-category response when a concrete event has low similarity to its abstract category text, Fig. 4 compares the ground-truth category probabilities of Baseline and Baseline + GapSeer in three UCF-Crime videos. Across the shaded low-similarity intervals, Baseline + GapSeer assigns higher probabilities to the ground-truth category for assault, arson, and burglary. This shows that GapSeer maintains a strong response to the valid category even where direct visual-text similarity is weak, supporting the use of category-specific semantic progression to represent concrete event instances.

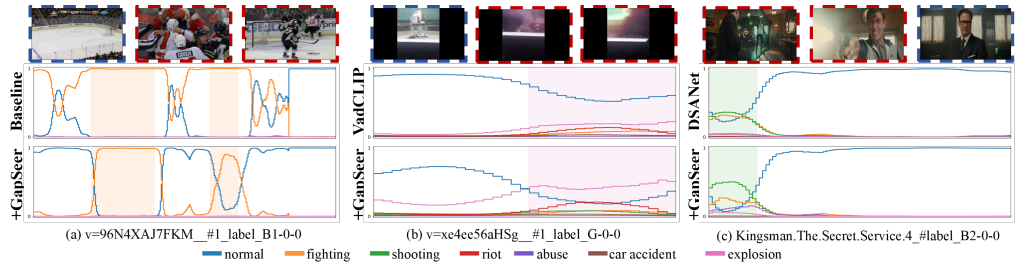

Figure 3: Fine-grained prediction curves comparing each host model with its GapSeer variant. Shaded regions indicate ground-truth anomalous intervals.

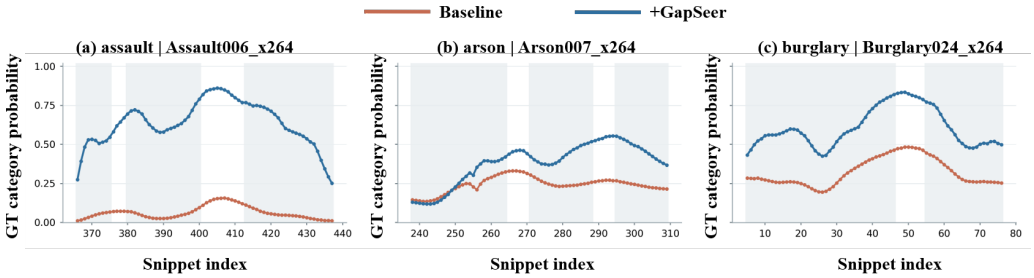

Figure 4: **Category responses under low visual-text similarity.** Ground-truth category probabilities for 72 consecutive ground-truth-overlapping anomaly snippets in each of three UCF-Crime videos. Shaded regions indicate snippets whose Baseline cosine similarity to the ground-truth category text is below the median of such similarities across ground-truth anomaly snippets in the UCF-Crime test set.

Distinguishing Visually Similar Anomalies. To examine whether GapSeer separates visually similar events with different category semantics, Fig. 5 compares category responses for three UCF-Crime snippet pairs. Despite the high visual similarity within each pair, Baseline assigns one snippet a higher probability for the other snippet's category than for its ground-truth category. Baseline + GapSeer reverses this preference in all three pairs while retaining a strong correct-category response for the paired snippet. For the assault snippet in Fig. 5(c), the assault probability increases from 0.02 to 0.58, while the shoplifting probability decreases from 0.60 to 0.02. This correction combines stronger evidence for the true category with reduced confusion from the competing category. These results show how category-specific semantic progression helps distinguish events that share visual content but belong to different anomaly categories.

Analysis across Visual Typicality Levels. To assess whether GapSeer maintains category recognition across different visual realizations of the same anomaly, Fig. 6(a) compares GT category Top-1 accuracy at three levels of within-class visual typicality. Baseline accuracy varies with visual typicality, although snippets in all three groups belong to their respective ground-truth categories. Baseline + GapSeer improves accuracy in every group: from 23.48% to 30.56% for Atypical snippets, from 27.85% to 28.07% for Medium snippets, and from 35.82% to 39.83% for Typical snippets. The largest improvement occurs in the Atypical group, suggesting that category-specific responses are particularly useful for instances that deviate from common within-class appearances. The smaller change in the Medium group also indicates that the benefit varies across typicality levels. The improvement across all three levels shows that category-specific semantic progression helps retain correct-category evidence across varied visual instances of the same anomaly.

Analysis of Category Separation. To examine whether the resulting category responses favor the ground-truth anomaly over competing anomaly categories, Fig. 6(b) compares the GT-competitor margins on XD-Violence. Baseline + GapSeer has a higher margin-exceedance curve across the displayed thresholds, increases the class-macro mean margin from 0.538 to 0.586, and yields a larger paired margin for 72.6% of GT snippets on a class-averaged basis. A larger margin means that the ground-truth response stands further above the strongest incorrect anomaly response. The threshold-

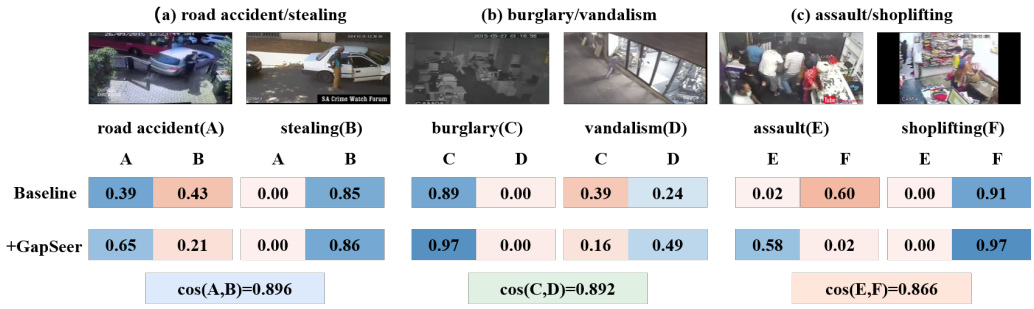

Figure 5: **Category responses for visually similar anomalies.** Three UCF-Crime snippet pairs from different ground-truth categories: (a) road accident and stealing, (b) burglary and vandalism, and (c) assault and shoplifting. Each cell gives the probability assigned to one of the two displayed categories by Baseline or Baseline + GapSeer. The cosine value beneath each pair measures similarity between their frozen CLIP visual features.

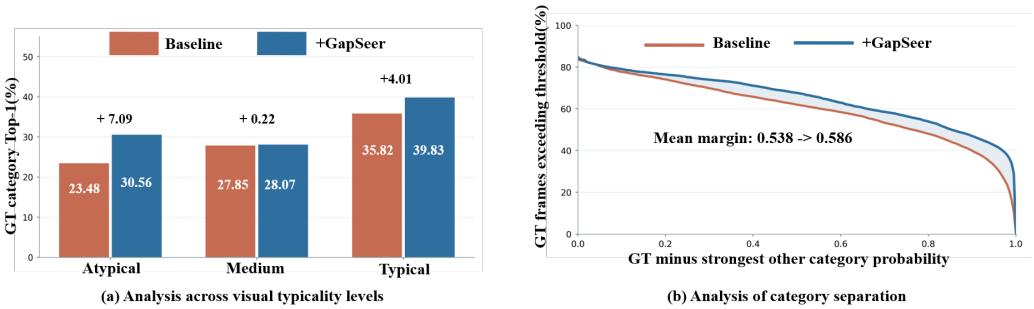

Figure 6: **Visual typicality and category separation.** (a) Ground-truth category Top-1 accuracy on UCF-Crime after grouping all ground-truth-overlapping snippets by within-class visual typicality. The groups are defined using the mean cosine similarity to the five nearest same-class snippets from other test videos and are fixed for both models. (b) Class-averaged share of ground-truth-overlapping snippets on XD-Violence whose ground-truth probability exceeds the strongest incorrect anomaly-category probability by at least the indicated threshold. The strongest incorrect category is selected separately for each model; normal is included in the softmax normalization but excluded from the competing anomaly categories.

wise and paired improvements thus show that category-specific semantic progression strengthens the category preference needed for fine-grained anomaly recognition.

5 Conclusion

In this paper, we investigate the challenge of modeling instance-to-category relations in weakly supervised fine-grained video anomaly detection. Existing methods typically derive category responses from overall visual-text similarity, which overlooks the semantic granularity gap between abstract category descriptions and diverse concrete event instances. To address this issue, we propose GapSeer, which models category-specific semantic development through the Hyperbolic Semantic Axis (HSA) and quantifies instance progression along each axis with the Signed Axial Response (SAR). By characterizing how visual instances realize different category semantics rather than measuring their overall similarity to category texts, GapSeer accommodates intra-class variation while reducing ambiguity among visually similar anomaly categories. Experiments on two benchmarks demonstrate the effectiveness and generality of the proposed semantic-axis response for fine-grained video anomaly detection.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing. We did not use generative AI tools for other tasks requiring disclosure, and the remaining disclosure tasks are not applicable to this work. We reviewed all AI-assisted text and independently revised it where necessary. We take responsibility for the final content of this work, including all text, claims, and other materials produced with the aid of generative AI.

REFERENCES

Junxi Chen, Liang Li, Li Su, Zheng-jun Zha, and Qingming Huang. Prompt-enhanced multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18319–18329, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_Prompt-Enhanced_Multiple_Instance_Learning_for_Weakly_Supervised_Video_Anomaly_Detection_CVPR_2024_paper.html.

Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. MGFN: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):387–395, 2023. doi: 10.1609/aaai.v37i1.25112.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7694–7731. PMLR, 2023.

Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. ReFLIP-VAD: Towards weakly supervised video anomaly detection via vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. doi: 10.1109/TCSVT.2024.3482007.

Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. MIST: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14009–14018, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Feng_MIST_Multiple_Instance_Self-Training_Framework_for_Video_Anomaly_Detection_CVPR_2021_paper.html.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1646–1655. PMLR, 2018.

Chao Huang, Yushu Shi, Jie Wen, Wei Wang, Yong Xu, and Xiaochun Cao. Ex-VAD: Explainable fine-grained video anomaly detection based on visual-language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25750–25761. PMLR, 2025.

Tianshan Liu, Kin-Man Lam, and Bing-Kun Bao. Injecting text clues for improving anomalous event detection from weakly labeled videos. *IEEE Transactions on Image Processing*, 33:5907–5920, 2024. doi: 10.1109/TIP.2024.3477351.

Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G. M. Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8022–8031, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lv_Unbiased_Multiple_Instance_Learning_for_Weakly_Supervised_Video_Anomaly_Detection_CVPR_2023_paper.html.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3779–3788. PMLR, 2018. URL https://proceedings.mlr.press/v80/nickel18a.html.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27263–27272, 2024. doi: 10.1109/CVPR52733.2024.02574.

Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 154–171, 2018.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4975–4986, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Tian_Weakly-Supervised_Video_Anomaly_Detection_With_Robust_Temporal_Feature_Magnitude_Learning_ICCV_2021_paper.html.

Yu Wang and Shiwei Chen. Learning event completeness for weakly supervised video anomaly detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62505–62517. PMLR, 2025.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–339, 2020.

Peng Wu, Xiaotao Liu, and Jing Liu. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia*, 25:1674–1685, 2023. doi: 10.1109/TMM.2022.3147369.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):6074–6082, 2024. doi: 10.1609/aaai.v38i6.28423.

Wei Wu, Xiaomeng Fan, Yuwei Wu, Zhi Gao, Pengxiang Li, Yunde Jia, and Mehrtash Harandi. Modality alignment across trees on heterogeneous hyperbolic manifolds. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=F1uJKsaf0M.

Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18899–18908, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Text_Prompt_with_Normality_Guidance_for_Weakly_Supervised_Video_Anomaly_Detection_CVPR_2024_paper.html.

Wenti Yin, Huaxin Zhang, Xiang Wang, Yuqing Lu, Yicheng Zhang, Bingquan Gong, Jialong Zuo, Li Yu, Changxin Gao, and Nong Sang. Learning to tell apart: Weakly supervised video anomaly detection via disentangled semantic alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):12027–12035, 2026.

Daiki Yoshikawa and Takashi Matsubara. PHyCLIP: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=I3Ct1eDmVI.

Tao Yu, Toni J. B. Liu, Albert Tseng, and Christopher De Sa. Shadow cones: A generalized framework for partial order embeddings. In *The Twelfth International Conference on Learning Representations*, 2024.

Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, and Ming-Hsuan Yang. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16271–16280, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Exploiting_Completeness_and_Uncertainty_of_Pseudo_Labels_for_Weakly_Supervised_CVPR_2023_paper.html.

Menghao Zhang, Yiyan Zhu, Pengfei Ren, Haifeng Sun, Qi Qi, Zirui Zhuang, Huazheng Wang, Lei Zhang, Jianxin Liao, and Jingyu Wang. Fine-vad: Towards fine-grained video anomaly detection via progressive cross-granularity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35514–35523, June 2026.

Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhong_Graph_Convolutional_Label_Noise_Cleaner_Train_a_Plug-And-Play_Action_Classifier_CVPR_2019_paper.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16816–16825, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhou_Conditional_Prompt_Learning_for_Vision-Language_Models_CVPR_2022_paper.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b. doi: 10.1007/s11263-022-01653-1.

A LORENTZ GEOMETRY

This appendix summarizes the hyperbolic geometry underlying HSA and SAR. We first introduce the Lorentz model and its geometric operations, then derive the signed projection onto a semantic axis.

A.1 LORENTZ MANIFOLD AND DISTANCE

Lorentzian inner product. We adopt the Lorentz model of d-dimensional hyperbolic space with curvature $-1/k$, where $k > 0$. Let $\boldsymbol{x} = [x_0, \boldsymbol{x}_s]$ and $\boldsymbol{y} = [y_0, \boldsymbol{y}_s]$ be vectors in $\mathbb{R}^{d+1}$, with time-like coordinates $x_0, y_0 \in \mathbb{R}$ and space-like components $\boldsymbol{x}_s, \boldsymbol{y}_s \in \mathbb{R}^d$. Their Lorentzian inner product is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}} = -x_0 y_0 + \boldsymbol{x}_s^\top \boldsymbol{y}_s. \tag{9}$$

Lorentz manifold. The Lorentz manifold is defined as

$$\mathcal{L}_k^d = \left\{ \boldsymbol{x} \in \mathbb{R}^{d+1} \;\middle|\; \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{L}} = -k,\ x_0 > 0 \right\}, \tag{10}$$

where $x_0 > 0$ selects the upper sheet of the hyperboloid. Its origin is $\boldsymbol{o} = [\sqrt{k}, \mathbf{0}]$, where $\mathbf{0} \in \mathbb{R}^d$. For any $\boldsymbol{x} \in \mathcal{L}_k^d$, the time-like coordinate is determined by its space-like component:

$$x_0 = \sqrt{k + \|\boldsymbol{x}_s\|_2^2}. \tag{11}$$

Geodesic distance. For $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_k^d$, their geodesic distance is

$$d_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{k}\ \operatorname{arcosh}\left(-\frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{L}}}{k} \right). \tag{12}$$

The radial distance from the origin is therefore

$$d_{\mathcal{L}}(\boldsymbol{o}, \boldsymbol{x}) = \sqrt{k}\ \operatorname{arcosh}\left(\frac{x_0}{\sqrt{k}} \right) = \sqrt{k}\ \operatorname{arsinh}\left(\frac{\|\boldsymbol{x}_s\|_2}{\sqrt{k}} \right). \tag{13}$$

This distance measures position relative to the origin. It does not specify a category-dependent direction.

A.2 TANGENT SPACES AND EXPONENTIAL MAPS

Tangent space and norm. For $\boldsymbol{x} \in \mathcal{L}_k^d$, the tangent space at $\boldsymbol{x}$ is

$$T_{\boldsymbol{x}} \mathcal{L}_k^d = \left\{ \boldsymbol{u} \in \mathbb{R}^{d+1} \;\middle|\; \langle \boldsymbol{x}, \boldsymbol{u} \rangle_{\mathcal{L}} = 0 \right\}. \tag{14}$$

The Lorentzian inner product restricted to this tangent space is positive definite. Thus, for $\boldsymbol{u} \in T_{\boldsymbol{x}} \mathcal{L}_k^d$, its tangent norm is

$$\|\boldsymbol{u}\|_{\mathcal{L}} = \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\mathcal{L}}}. \tag{15}$$

At the origin, the tangent space takes the simple form

$$T_{\boldsymbol{o}} \mathcal{L}_k^d = \left\{ [0, \mathbf{z}] \;\middle|\; \mathbf{z} \in \mathbb{R}^d \right\}. \tag{16}$$

This identifies a d-dimensional encoder feature with a tangent vector by prepending a zero time-like coordinate.

Tangent-space projection. An ambient vector $\boldsymbol{u} \in \mathbb{R}^{d+1}$ can be projected onto $T_{\boldsymbol{x}} \mathcal{L}_k^d$ by

$$\operatorname{proj}_{\boldsymbol{x}}^k(\boldsymbol{u}) = \boldsymbol{u} + \frac{\langle \boldsymbol{x}, \boldsymbol{u} \rangle_{\mathcal{L}}}{k} \boldsymbol{x}. \tag{17}$$

The resulting vector satisfies $\langle \boldsymbol{x}, \operatorname{proj}_{\boldsymbol{x}}^k(\boldsymbol{u}) \rangle_{\mathcal{L}} = 0$. This operation projects an ambient vector onto a tangent space; projection of a manifold point onto a geodesic is defined separately below.

Exponential map. For a nonzero tangent vector $\boldsymbol{u} \in T_{\boldsymbol{x}} \mathcal{L}_k^d$, the exponential map is

$$\exp_{\boldsymbol{x}}^k(\boldsymbol{u}) = \cosh\left(\frac{\|\boldsymbol{u}\|_{\mathcal{L}}}{\sqrt{k}} \right) \boldsymbol{x} + \sqrt{k}\ \sinh\left(\frac{\|\boldsymbol{u}\|_{\mathcal{L}}}{\sqrt{k}} \right) \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_{\mathcal{L}}}. \tag{18}$$

At the zero tangent vector, the map is defined by continuity as $\exp_{\boldsymbol{x}}^k(\mathbf{0}_{d+1}) = \boldsymbol{x}$. At the origin, a nonzero feature $\mathbf{z} \in \mathbb{R}^d$ is mapped as

$$\exp_{\boldsymbol{o}}^k([0, \mathbf{z}]) = \left[\sqrt{k}\cosh\left(\frac{\|\mathbf{z}\|_2}{\sqrt{k}}\right),\ \sqrt{k}\sinh\left(\frac{\|\mathbf{z}\|_2}{\sqrt{k}}\right)\frac{\mathbf{z}}{\|\mathbf{z}\|_2}\right]. \tag{19}$$

Its radial distance is $d_{\mathcal{L}}(\boldsymbol{o}, \exp_{\boldsymbol{o}}^k([0, \mathbf{z}])) = \|\mathbf{z}\|_2$. Consequently, the modality-specific scaling factors in Eq. 1 control radial distances before alignment.

A.3 Oriented Geodesics and Angles

Unit-speed geodesics. For $\boldsymbol{x} \in \mathcal{L}_k^d$ and a unit tangent vector $\mathbf{a} \in T_{\boldsymbol{x}}\mathcal{L}_k^d$, the corresponding geodesic is

$$\gamma(u) = \exp_{\boldsymbol{x}}^k(u\mathbf{a}) = \cosh\left(\frac{u}{\sqrt{k}}\right)\boldsymbol{x} + \sqrt{k}\,\sinh\left(\frac{u}{\sqrt{k}}\right)\mathbf{a}, \qquad u \in \mathbb{R}. \tag{20}$$

This parameterization satisfies $\gamma(0) = \boldsymbol{x}$, $\gamma'(0) = \mathbf{a}$, and

$$d_{\mathcal{L}}\big(\gamma(u_1), \gamma(u_2)\big) = |u_1 - u_2|. \tag{21}$$

Thus, u is signed arc length from $\boldsymbol{x}$. The domain $u \in \mathbb{R}$ defines a complete oriented geodesic, whereas $u \geq 0$ defines its forward ray.

Radial orientation of HSA. For a category representation $\mathbf{t}_c \neq \boldsymbol{o}$, the vector $\mathrm{proj}_{\mathbf{t}_c}^k(\boldsymbol{o})$ points toward the origin. Its negative normalized direction defines the outward unit tangent $\mathbf{a}_c$ in Eq. 2. Writing $r_c = d_{\mathcal{L}}(\boldsymbol{o}, \mathbf{t}_c)$, the same direction can be expressed as

$$\mathbf{a}_c = \frac{\cosh(r_c/\sqrt{k})\mathbf{t}_c - \boldsymbol{o}}{\sqrt{k}\sinh(r_c/\sqrt{k})}. \tag{22}$$

The resulting axis $\gamma_c(u) = \exp_{\mathbf{t}_c}^k(u\mathbf{a}_c)$ passes through the origin at $u = -r_c$. Its positive ray starts at $\mathbf{t}_c$ and extends away from the origin. At $\mathbf{t}_c = \boldsymbol{o}$, a category-specific radial direction is undefined.

Geodesic directions and angles. For distinct points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_k^d$, let $\delta = d_{\mathcal{L}}(\boldsymbol{x}, \boldsymbol{y})$. The unit tangent at $\boldsymbol{x}$ pointing toward $\boldsymbol{y}$ is

$$\mathbf{b}_{\boldsymbol{x}\to\boldsymbol{y}} = \frac{\mathrm{proj}_{\boldsymbol{x}}^k(\boldsymbol{y})}{\|\mathrm{proj}_{\boldsymbol{x}}^k(\boldsymbol{y})\|_{\mathcal{L}}} = \frac{\boldsymbol{y} - \cosh(\delta/\sqrt{k})\boldsymbol{x}}{\sqrt{k}\sinh(\delta/\sqrt{k})}. \tag{23}$$

For two nonzero tangent vectors $\boldsymbol{u}, \mathbf{w} \in T_{\boldsymbol{x}}\mathcal{L}_k^d$, the angle $\theta \in [0, \pi]$ between them satisfies

$$\cos\theta = \frac{\langle \boldsymbol{u}, \mathbf{w}\rangle_{\mathcal{L}}}{\|\boldsymbol{u}\|_{\mathcal{L}}\|\mathbf{w}\|_{\mathcal{L}}}. \tag{24}$$

In SAR, these directions are the outward axis tangent $\mathbf{a}_c$ and the unit tangent from $\mathbf{t}_c$ toward $\mathbf{h}_t$, yielding Eq. 4.

A.4 Hyperbolic Triangle Relations

Hyperbolic law of cosines. Consider a geodesic triangle with side lengths ℓ_1, ℓ_2, ℓ_3. Let θ be the angle between the sides of lengths ℓ_1 and ℓ_2, opposite the side of length ℓ_3. For curvature $-1/k$,

$$\cosh\left(\frac{\ell_3}{\sqrt{k}}\right) = \cosh\left(\frac{\ell_1}{\sqrt{k}}\right)\cosh\left(\frac{\ell_2}{\sqrt{k}}\right) - \sinh\left(\frac{\ell_1}{\sqrt{k}}\right)\sinh\left(\frac{\ell_2}{\sqrt{k}}\right)\cos\theta. \tag{25}$$

Right-triangle identities. For a right triangle with perpendicular legs of lengths ℓ_1 and ℓ_2 and hypotenuse of length ℓ_3, the cosine law reduces to

$$\cosh\left(\frac{\ell_3}{\sqrt{k}}\right) = \cosh\left(\frac{\ell_1}{\sqrt{k}}\right)\cosh\left(\frac{\ell_2}{\sqrt{k}}\right). \tag{26}$$

Let β denote the angle between the leg of length ℓ_1 and the hypotenuse. Applying the cosine law at this vertex and substituting the preceding identity gives

$$\tanh\left(\frac{\ell_1}{\sqrt{k}}\right) = \tanh\left(\frac{\ell_3}{\sqrt{k}}\right)\cos\beta. \tag{27}$$

This relation connects the hypotenuse and its direction to the length of the adjacent leg. The signed-axis formulation below extends it to projections on either side of the category representation.

A.5 Geodesic Projection and Signed Axial Response

Projection onto the semantic axis. For a snippet representation $\mathbf{h}_t$, its closest point on the complete axis γ_c is denoted by $\mathbf{p}_{c,t}$. Since γ_c has unit speed, the signed coordinate of this point is

$$R_{c,t} = \arg\min_{u \in \mathbb{R}} d_{\mathcal{L}}\big(\mathbf{h}_t, \gamma_c(u)\big), \qquad \mathbf{p}_{c,t} = \gamma_c(R_{c,t}). \tag{28}$$

When $\mathbf{h}_t$ lies off the axis, the geodesic joining $\mathbf{h}_t$ to $\mathbf{p}_{c,t}$ is perpendicular to the axis. The projection is onto the complete geodesic, so its coordinate can be positive, zero, or negative.

Derivation of the signed coordinate. For $d_{c,t} = d_{\mathcal{L}}(\mathbf{t}_c, \mathbf{h}_t) > 0$, let $\phi_{c,t}$ be the angle between $\mathbf{a}_c$ and the geodesic direction toward $\mathbf{h}_t$. Substituting the axis parameterization into the Lorentz distance formula yields

$$\cosh\left(\frac{d_{\mathcal{L}}(\mathbf{h}_t, \gamma_c(u))}{\sqrt{k}}\right) = \cosh\left(\frac{d_{c,t}}{\sqrt{k}}\right)\cosh\left(\frac{u}{\sqrt{k}}\right) - \sinh\left(\frac{d_{c,t}}{\sqrt{k}}\right)\sinh\left(\frac{u}{\sqrt{k}}\right)\cos\phi_{c,t}. \tag{29}$$

Because $\cosh$ is strictly increasing on nonnegative arguments, minimizing the distance is equivalent to minimizing the right-hand side. Setting its derivative with respect to u to zero gives

$$\tanh\left(\frac{R_{c,t}}{\sqrt{k}}\right) = \tanh\left(\frac{d_{c,t}}{\sqrt{k}}\right)\cos\phi_{c,t}. \tag{30}$$

The right-hand side has magnitude less than one. The distance objective has a unique minimum, giving

$$R_{c,t} = \sqrt{k}\ \operatorname{artanh}\left[\tanh\left(\frac{d_{c,t}}{\sqrt{k}}\right)\cos\phi_{c,t}\right], \qquad d_{c,t} > 0. \tag{31}$$

When $\mathbf{h}_t = \mathbf{t}_c$, the closest point is $\mathbf{t}_c$ and $R_{c,t} = 0$, recovering the full definition in Eq. 6.

An equivalent expression follows directly from Lorentzian inner products:

$$R_{c,t} = \sqrt{k}\ \operatorname{artanh}\left(\frac{\sqrt{k}\,\langle \mathbf{a}_c, \mathbf{h}_t\rangle_{\mathcal{L}}}{-\langle \mathbf{t}_c, \mathbf{h}_t\rangle_{\mathcal{L}}}\right). \tag{32}$$

This expression also returns zero when $\mathbf{h}_t = \mathbf{t}_c$.

Axial displacement and orthogonal deviation. Let $\rho_{c,t} = d_{\mathcal{L}}(\mathbf{h}_t, \mathbf{p}_{c,t})$ be the orthogonal distance from the snippet to the axis. The right-triangle identity gives

$$\cosh\left(\frac{d_{c,t}}{\sqrt{k}}\right) = \cosh\left(\frac{R_{c,t}}{\sqrt{k}}\right)\cosh\left(\frac{\rho_{c,t}}{\sqrt{k}}\right). \tag{33}$$

Because $\cosh$ is even, this relation holds for either sign of $R_{c,t}$. The two quantities describe different aspects of the snippet's position: $R_{c,t}$ measures signed displacement along the axis, whereas $\rho_{c,t}$ measures distance from it. SAR uses the signed axial coordinate as the category response.

Sign and magnitude. For $d_{c,t} > 0$, the signed coordinate satisfies

$$|R_{c,t}| \le d_{c,t}, \qquad \operatorname{sgn}(R_{c,t}) = \operatorname{sgn}(\cos\phi_{c,t}). \tag{34}$$

A snippet on the positive axis has $R_{c,t} = d_{c,t}$; one on the negative axis has $R_{c,t} = -d_{c,t}$. A perpendicular direction gives $R_{c,t} = 0$, even when the snippet is distant from the category representation.

Local Euclidean behavior. For $d_{c,t}/\sqrt{k} \ll 1$, expanding the hyperbolic functions gives

$$R_{c,t} = d_{c,t}\cos\phi_{c,t} + O\left(\frac{d_{c,t}^3}{k}\right). \tag{35}$$

Thus, at distances small relative to the curvature scale, SAR approaches the Euclidean signed projection. At larger distances, the hyperbolic triangle relation determines the axial coordinate.

B Integration with MIL Supervision

We detail the supervision of GapSeer in the Fine-VAD-based framework (Zhang et al., 2026) used in our main experiments. The coarse stage distinguishes normal and anomalous snippets, whereas the fine stage considers all C categories, including normal. At either stage $s \in \{\text{coarse}, \text{fine}\}$, the alignment logits are

$$M_{c,t}^{(s)} = R_{c,t}^{(s)} / \tau_s, \tag{36}$$

where $R_{c,t}^{(s)}$ is the SAR defined in Eq. 6 and $\tau_s > 0$ is the stage-specific temperature. All temporal aggregation uses valid snippets only. For a video containing T valid snippets, we select $K = \lfloor T/D \rfloor + 1$ snippets, where D is the pooling divisor.

Coarse supervision. We first apply softmax over the normal and anomaly logits at each snippet. With indices 0 and 1 denoting normal and anomaly, respectively, the snippet anomaly probability is

$$q_t = \frac{\exp(M_{1,t}^{(\text{coarse})})}{\sum_{j=0}^{1} \exp(M_{j,t}^{(\text{coarse})})}. \tag{37}$$

Let $\mathcal{I}^b$ contain the indices of the K largest values of q_t. The video-level anomaly probability and binary loss are

$$\hat{y}^b = \frac{1}{K} \sum_{t \in \mathcal{I}^b} q_t, \qquad \mathcal{L}_{\text{bce}} = -y^b \log \hat{y}^b - (1 - y^b) \log(1 - \hat{y}^b). \tag{38}$$

Fine supervision. For each category c, let $\mathcal{I}_c$ contain the indices of its K largest fine-stage logits. We pool these logits before applying softmax across categories:

$$z_c = \frac{1}{K} \sum_{t \in \mathcal{I}_c} M_{c,t}^{(\text{fine})}, \qquad \hat{p}_c = \frac{\exp(z_c)}{\sum_{j=0}^{C-1} \exp(z_j)}. \tag{39}$$

The category loss uses the video-level multi-hot label:

$$\mathcal{L}_{\text{ce}} = -\sum_{c=0}^{C-1} y_c^m \log \hat{p}_c. \tag{40}$$

We sum the loss contributions of all positive categories without normalizing the multi-hot target. For a single positive category, this reduces to the usual multiclass cross-entropy.

The training objective averages $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{ce}}$ over the videos in each batch.

C Additional Experimental Analysis

We provide additional evidence for the category-response design of GapSeer. The following experiments examine multi-category cone membership, class-wise category responses, and category separation across different levels of Baseline ambiguity.

C.1 Complete Ablation Study of Response Functions

Table 4 reports the complete response-function ablation on both benchmarks. Consistent with Sec. 4.4, Baseline + GapSeer achieves the highest mAP at every evaluated IoU threshold on XD-Violence and UCF-Crime.

C.2 Multi-Category Entailment-Cone Membership

To examine the ambiguity of cone inclusion described in the Introduction, we evaluated every anomaly-category cone for all abnormal GT-overlapping test snippets. Using the representations of the Entailment-cone response model, we determined membership in category cone c by

$$\psi(\mathbf{t}_c) - \angle_O(\mathbf{t}_c, \mathbf{h}_t) \geq 0,$$

Table 4: **Complete ablation study of response functions.** Temporal detection mAP (%) at IoU thresholds from 0.1 to 0.5. AVG denotes the average across these thresholds. Baseline uses cosine similarity; Baseline + GapSeer uses SAR.

Response function	XD-Violence						UCF-Crime					
	0.1	0.2	0.3	0.4	0.5	AVG	0.1	0.2	0.3	0.4	0.5	AVG
Baseline	45.51	36.19	30.96	26.23	22.50	32.28	15.41	12.28	11.43	10.79	9.64	11.91
Angular response	37.06	31.57	27.47	23.66	21.38	28.23	6.55	6.54	6.54	6.51	6.39	6.51
Axis-distance response	36.34	31.05	26.36	22.92	21.25	27.58	6.54	6.53	6.49	6.45	6.29	6.46
Lorentz-distance response	40.32	32.31	26.80	23.38	20.73	28.71	6.75	6.75	6.75	6.71	6.67	6.73
Entailment-cone response	37.45	30.60	26.36	22.95	21.29	27.73	6.55	6.55	6.54	6.51	6.39	6.51
Radial-distance-difference response	35.68	29.19	25.04	21.88	20.35	26.43	11.81	11.43	10.56	9.14	9.12	10.41
Euclidean axial response	43.10	35.51	29.94	25.95	22.80	31.46	19.06	15.25	14.20	12.85	11.72	14.62
Baseline + GapSeer	**45.82**	**38.90**	**33.66**	**28.01**	**23.76**	**34.03**	**23.80**	**16.85**	**14.92**	**13.64**	**12.80**	**16.40**

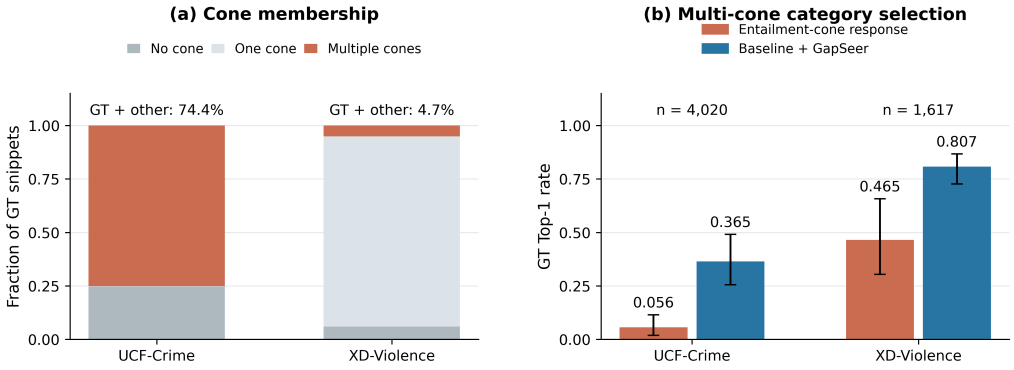

Figure 7: **Multi-category entailment-cone membership.** (a) Distribution of abnormal GT-overlapping snippets according to the number of anomaly-category cones containing them. Annotations report the fraction inside both the GT cone and at least one incorrect cone. (b) GT Top-1 accuracy among anomaly categories on this fixed subset, comparing Entailment-cone response with Baseline + GapSeer. Error bars indicate 95% confidence intervals from 2,000 video-cluster bootstrap resamples.

where ψ denotes the entailment aperture and $\angle_O$ denotes the entailment angle. The multi-cone subset contains snippets inside both their GT cone and at least one incorrect anomaly-category cone.

Fig. 7(a) shows that 4,020 of 5,400 snippets on UCF-Crime (74.44%) and 1,617 of 34,671 snippets on XD-Violence (4.66%) belonged to both their GT cone and an incorrect cone. For these snippets, the inclusion condition accepts multiple category hypotheses for the same event. These results establish that cone inclusion can identify semantic compatibility with several anomaly categories simultaneously, supporting the distinction between compatibility and category discrimination made in the Introduction.

We then kept this subset fixed and compared the category rankings produced by Entailment-cone response and Baseline + GapSeer. In Fig. 7(b), GT Top-1 accuracy increased from 5.65% to 36.49% on UCF-Crime and from 46.51% to 80.71% on XD-Violence. Incorrect-to-correct transitions accounted for 31.57% and 41.31% of the respective subsets, compared with correct-to-incorrect transitions of 0.72% and 7.11%. These results show that SAR improves GT-category selection among snippets belonging to multiple anomaly-category cones, supporting the use of category-specific semantic progression to resolve ambiguous cone membership.

The cone analysis counts each temporal snippet once. For snippets overlapping multiple category annotations, the GT category follows the last overlapping annotation. The class-wise response and ambiguity analyses retain individual snippet–GT-category assignments. Accordingly, XD-Violence contributes 34,671 unique snippets to the cone analysis and 34,753 assignments to the response analyses.

Fig. 8 provides three visual examples of this ambiguity. The snippet pairs have visual cosine similarities of 0.928, 0.896, and 0.892. In each pair, both snippets lie inside both displayed category cones,

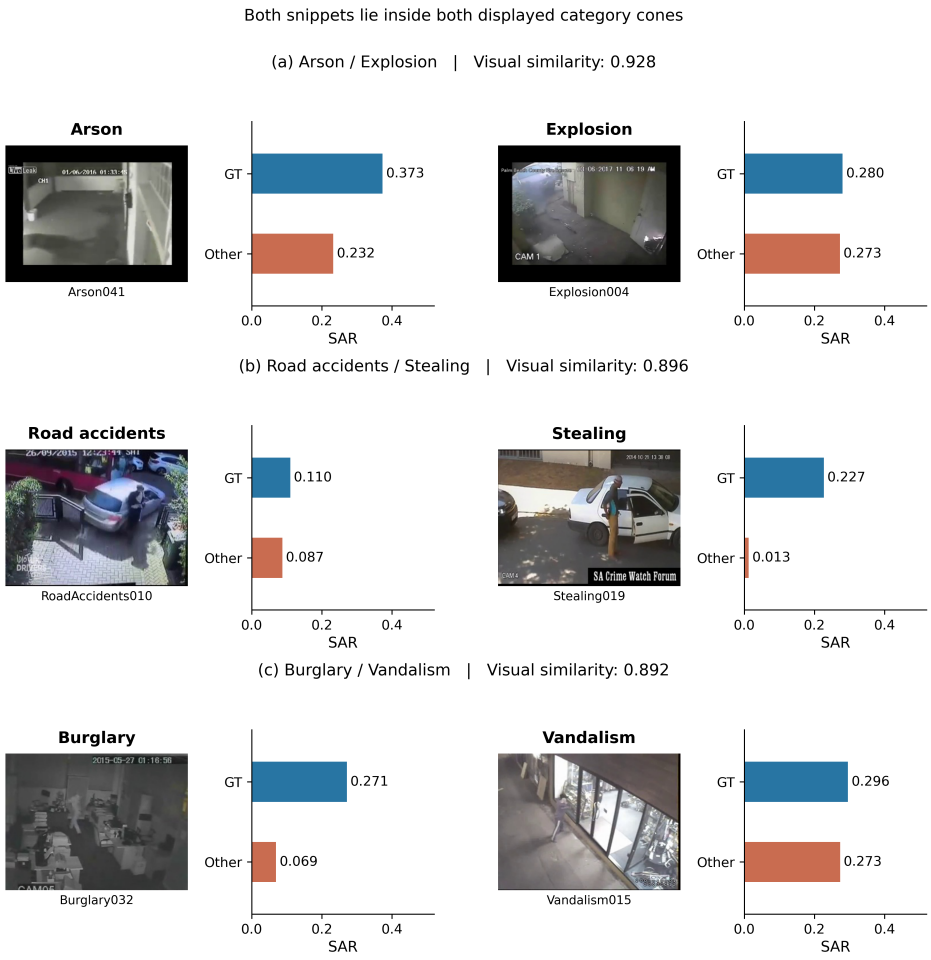

Figure 8: **Visually similar events with multiple cone memberships.** Three UCF-Crime snippet pairs from different videos and GT categories: arson and explosion, road accidents and stealing, and burglary and vandalism. Visual cosine similarity is computed from frozen CLIP features. Cone indicators use the Entailment-cone response model; bars show SAR values from Baseline + GapSeer for the two displayed categories.

despite having different GT categories. Baseline + GapSeer assigns each snippet a higher SAR for its own GT category than for the paired category. These examples show that visually similar events can share the same binary cone memberships while receiving higher SAR values along their respective GT-category HSAs. Thus, signed axial progression provides category discrimination beyond cone inclusion.

C.3 Class-Wise Category Responses

To examine category separation across anomaly classes, we compared the GT probability with the strongest incorrect anomaly-category probability for every class. For each snippet and model, the strongest incorrect category was selected among all anomaly categories other than GT. Probabilities were normalized over anomaly categories, excluding normal, and class means were weighted by the number of GT-overlapping frames.

Fig. 9 and Fig. 10 show the class-wise changes in GT and competing responses. For shooting on XD-Violence, the mean GT probability increased from 0.602 to 0.699, while the mean strongest incorrect probability decreased from 0.312 to 0.240. The resulting margin increased from 0.291 to 0.459. For road accidents on UCF-Crime, the corresponding margin increased from 0.290 to 0.604. These

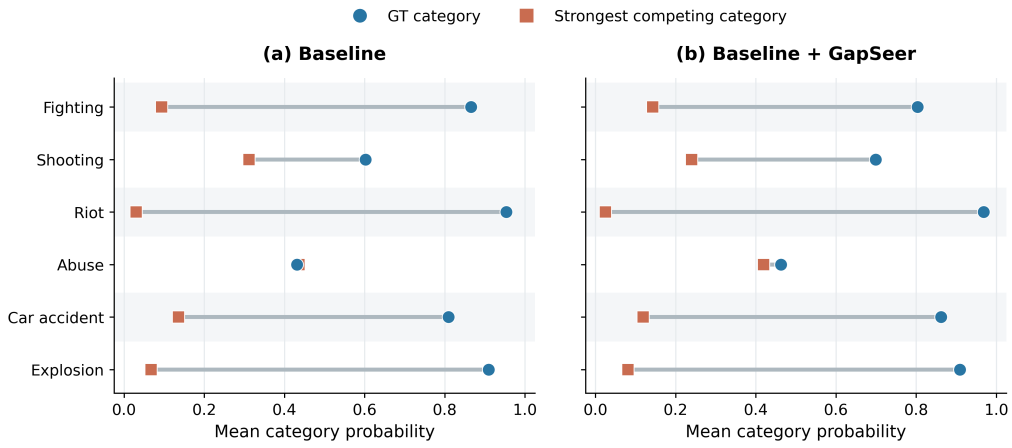

Figure 9: **Class-wise category responses on XD-Violence.** Circles show mean GT probabilities, and squares show mean strongest incorrect anomaly-category probabilities for (a) Baseline and (b) Baseline + GapSeer. Means are weighted by GT-overlapping frames. Probabilities are normalized over anomaly categories, excluding normal. Subtracting the square's horizontal coordinate from the circle's gives the mean GT–competitor margin.

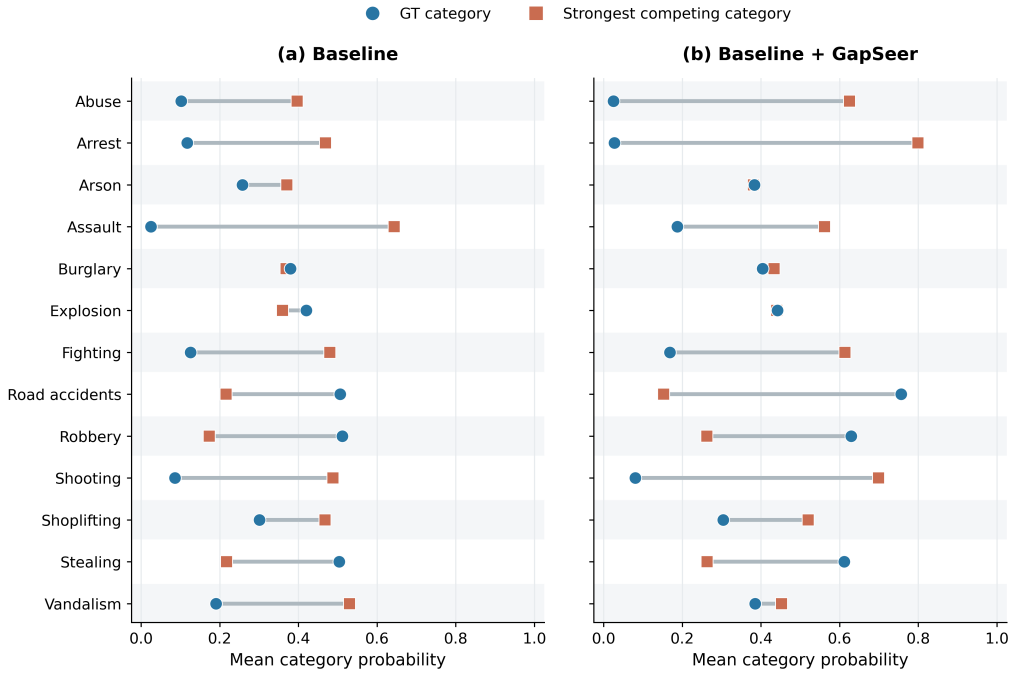

Figure 10: **Class-wise category responses on UCF-Crime.** Circles show mean GT probabilities, and squares show mean strongest incorrect anomaly-category probabilities for (a) Baseline and (b) Baseline + GapSeer. Means are weighted by GT-overlapping frames. Probabilities are normalized over anomaly categories, excluding normal. Subtracting the square's horizontal coordinate from the circle's gives the mean GT–competitor margin.

results show that the separation gains in these classes arise from both increased GT probabilities and reduced strongest-competitor probabilities. This supports the role of SAR in strengthening the relative category evidence required for fine-grained anomaly recognition.

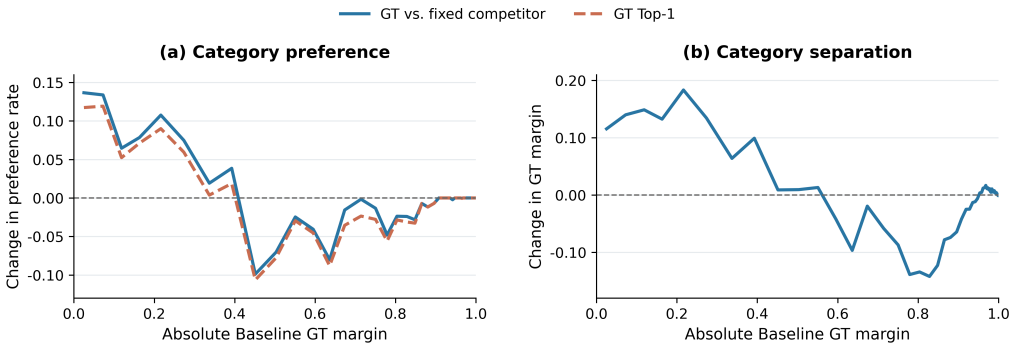

Figure 11: **Category preference and separation across Baseline ambiguity levels.** Each point summarizes one of 80 groups of GT snippet–category assignments on XD-Violence. Its horizontal coordinate is the group's median absolute Baseline GT–competitor margin. (a) Changes in the fraction of assignments favoring GT over the fixed Baseline competitor and in GT Top-1 accuracy among anomaly categories. (b) Change in the mean GT–competitor margin, selecting the strongest incorrect anomaly category separately for each model. All changes are Baseline + GapSeer minus Baseline.

The relative change in the two responses also accounts for the class-wise variation visible in the plots. For burglary on UCF-Crime, the GT probability increased by 0.025, while the strongest incorrect probability increased by 0.065; the mean margin therefore decreased by 0.041. Thus, class-wise separation improves when the change in GT probability exceeds the change in the strongest incorrect probability. The paired response measurements show both how GT evidence is strengthened and how that evidence translates into discrimination against competing anomaly categories.

C.4 Category Separation Across Baseline Ambiguity Levels

To examine the effect of GapSeer when competing categories receive similar Baseline responses, we sorted all 34,753 GT snippet–category assignments on XD-Violence by the absolute Baseline GT–competitor margin and divided them into 80 approximately equal-sized groups. A smaller absolute margin indicates closer GT and strongest incorrect responses. Group membership was fixed for both methods. Following the margin definition in Fig. 6(b), probabilities were normalized over all categories, including normal, while competing categories were restricted to incorrect anomaly categories.

Fig. 11(a) evaluates category preference against both the original Baseline competitor and the complete set of anomaly categories. The fixed-competitor measure tests whether the original category pair is better separated, whereas GT Top-1 tests whether the GT category ranks above every incorrect anomaly category. Fig. 11(b) measures the change in separation from each model's own strongest competitor.

In the group with the smallest absolute Baseline margin, the median was 0.025. GT preference over the fixed competitor increased by 0.137, GT Top-1 accuracy increased by 0.117, and the mean GT–competitor margin increased by 0.115. Improvement in all three measures shows that GapSeer strengthens both the preference over the original competitor and the GT ranking among all anomaly categories where Baseline responses are most ambiguous.

At the high-margin end, the final six groups had median absolute Baseline margins between 0.9993 and 0.9999. Both preference-rate changes were zero, while mean margin changes ranged from -0.0013 to -0.0006. Together, these results show that GapSeer improves GT-category ranking and separation in the most ambiguous Baseline group, while the high-margin groups retain unchanged aggregate preference rates. Thus, category-specific semantic progression strengthens discrimination where GT and competing category responses are close.