SphereVAD: Training-Free Video Anomaly Detection via Geodesic Inference on the Unit Hypersphere

Chao Huang[1] **Pengfei Wei**[1] **Wei Wang**[1] **Jie Wen**[2] **Zhihua Wang**[1,3]
Li Shen[1] **Wenqi Ren**[1] **Xiaochun Cao**[1*]

[1]Shenzhen Campus of Sun Yat-sen University
[2]Harbin Institute of Technology, Shenzhen [3]SunYield Technology Co., Ltd.

{huangch253, wangwei29, wangzh, shenli6, caoxiaochun}@email.sysu.edu.cn
weipf@mail2.sysu.edu.cn wenjie@hit.edu.cn

Abstract

Video anomaly detection (VAD) aims to automatically identify events that deviate from normal patterns in untrimmed surveillance videos. Existing methods universally depend on large-scale annotations or task-specific training procedures, severely limiting their rapid deployment to novel scenes. We observe that intermediate-layer features of pre-trained multimodal large language models (MLLMs) already encode rich anomaly semantics, yet existing approaches rely on the language output pathway and fail to exploit the geometric discriminability latent in these representations. Based on this finding, we propose **SphereVAD**, a fully training-free, zero-shot VAD framework that recasts anomaly discrimination as von Mises–Fisher (vMF) likelihood-ratio geodesic inference on the unit hypersphere, unleashing latent discriminability through principled geometric reasoning rather than learning new representations. Specifically, SphereVAD first applies Fréchet mean centering to unfold feature distributions and eliminate domain biases, then employs Holistic Scene Attention (HSA) to reinforce feature consistency using cross-video priors, and finally performs vMF-guided Spherical Geodesic Pulling (SGP) to align ambiguous segments with directional prototypes on the spherical manifold. This training-free pipeline requires only minimal synthetic images for calibration. SphereVAD establishes new state-of-the-art results among zero-shot, training-free methods on three major benchmarks and remains competitive with weakly supervised baselines. The code is publicly available at `https://github.com/S1inetzz/SphereVAD`.

1 Introduction

Video anomaly detection (VAD) aims to automatically identify events that deviate from expected behaviour in untrimmed surveillance videos, with broad applications in public safety, intelligent transportation, and beyond [37, 40]. The key challenge of this task lies in the open-ended nature of anomalies: anomalous event categories are diverse and virtually inexhaustible, rendering large-scale annotation prohibitively expensive. Existing approaches, whether weakly supervised [40, 43], unsupervised / self-trained [42, 34], or few-shot paradigms [29], all rely on some form of training procedure and therefore face bottlenecks including high computational cost, strong dependence on target-domain data, and the necessity of retraining when transferring across scenes.

Recent advances in vision-language models and multimodal large language models (MLLMs) [25, 36] have opened new possibilities for training-free VAD. Several methods attempt to leverage the language

*Corresponding author.

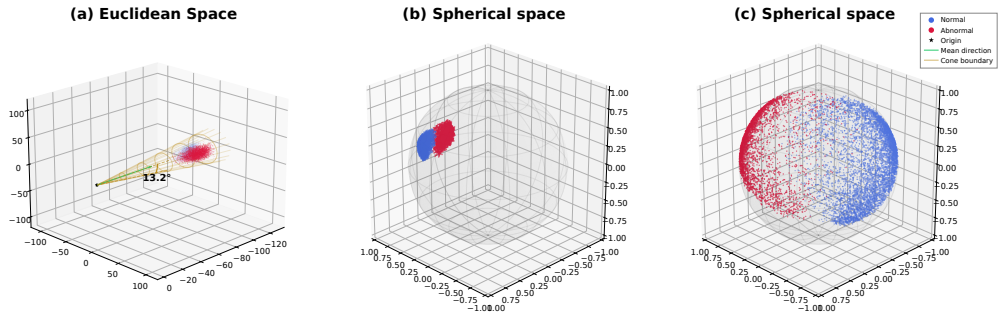


Figure 1: Motivation of SphereVAD on XD-Violence. Features are extracted from intermediate layers of a frozen Qwen3.5. (a) Original features in Euclidean space concentrate within a $\sim 13^\circ$ cone. (b) After ℓ_2-normalisation, features remain in a small spherical cap. (c) After spherical centering, features unfold over the full hypersphere with substantially improved class separation.

reasoning capability of MLLMs to judge anomalies directly [52], yet a significant performance gap with respect to supervised counterparts persists. We find that MLLM intermediate-layer features already encode rich discriminative information about anomalies. The real bottleneck is that current methods rely exclusively on the language output pathway for inference, leaving the abundant semantic representations in these intermediate layers largely untapped.

However, directly utilising MLLM intermediate-layer features for anomaly detection exposes three widely overlooked geometric issues. **First, conical collapse of features.** As illustrated in Figure 1(a), MLLM intermediate-layer features exhibit a highly conical distribution in Euclidean space, concentrated within a narrow cone of merely $\sim 13^\circ$. In such a crowded space, Euclidean distance metrics become extremely fragile, the most discriminative directional information remains underexploited. Even after ℓ_2-normalisation onto the unit hypersphere (Figure 1(b)), the distribution remains clustered in a small spherical cap (angular standard deviation $\approx 1.7^\circ$, as confirmed by the narrow spread in Figure 3, left column). Spherical centering (Figure 1(c)) effectively unfolds these features over the full hypersphere (angular standard deviation $\approx 17.1^\circ$, Figure 3, middle column), and the separation between normal and anomalous features improves significantly. **Second, domain shift across different data sources.** Features extracted from different data sources occupy distinct regions on the hypersphere, manifesting as a systematic rotational offset between their respective mean directions. If left unaligned, this rotational bias causes reference directions calibrated on one domain to be misaligned with features from another, leading to substantial performance degradation (see Appendix F.3 for empirical evidence). **Third, ambiguous features near the decision boundary.** In decision-space-based methods, a considerable proportion of features fall into the ambiguous zone between normal and anomalous regions. These borderline features are highly sensitive to small perturbations and can easily be assigned incorrect labels, undermining both the accuracy and robustness of the detection.

Based on the observations above, we propose SphereVAD, a fully training-free video anomaly detection framework that recasts anomaly discrimination as geodesic inference on the unit hypersphere $\mathcal{S}^{D-1}$. SphereVAD comprises three complementary components. **1. Unified Fréchet Mean Spherical Centering**, which aligns features from different data sources to a common tangent-space coordinate system while simultaneously unfolding the conical distribution over the entire sphere. **2. Holistic Scene Attention (HSA)**, which constructs cross-video attention maps based on scene-similarity priors to reinforce feature consistency. **3. vMF-guided Spherical Geodesic Pulling (SGP)**, which moves ambiguous segments along geodesics toward the dominant vMF prototype direction, with pull strength that adapts to the degree of uncertainty, ensuring that features always remain on the spherical manifold. The entire pipeline requires only approximately 2,000 synthetic calibration images, involves no gradient descent, and contains only four core hyperparameters.

Our contributions are summarised as follows:

- We propose SphereVAD, the first training-free framework that models video anomaly detection as vMF likelihood-ratio geodesic inference on the unit hypersphere, offering genuine plug-and-play capability.

- We propose a unified Fréchet mean spherical centering strategy that effectively alleviates the rotational bias across different data sources. We further prove that this operation does not reduce, and typically enlarges, inter-class angular distances (Proposition 1).
- We design HSA, a cross-video holistic scene attention module that strengthens feature consistency across videos, together with SGP, a score-adaptive spherical geodesic pulling mechanism that corrects the direction of ambiguous features under manifold constraints with pull strength adapting inversely to confidence.
- SphereVAD establishes state-of-the-art results among zero-shot, training-free approaches on XD-Violence (AP 86.99%, +16.83 % over PANDA), UCF-Crime (AUC 86.38%, +1.49 % over PANDA), and UBnormal (AUC 76.46%), substantially narrowing the gap to task-specific training paradigms and remaining competitive with weakly supervised baselines.

2 Related Work

Video Anomaly Detection. Traditional VAD approaches learn task-specific representations through self-supervised generative modeling [20] or prototype-guided prediction [19]. Weakly supervised methods, including MIL [40], Prompt-Enhanced MIL [8], and self-guided temporal discriminative transformers [16], learn anomaly discrimination from video-level labels. Despite reducing annotation requirements, these approaches still require task-specific training. Recent vision-language approaches advance semantic anomaly understanding [10, 9]: Ex-VAD [17] integrates visual information with textual explanations for fine-grained detection, while Vad-R1 [18] develops perception-to-cognition chain-of-thought reasoning through reinforcement learning. Zero-shot methods such as LAVAD [52], LAVIDA [12], and Sherlock [31] further explore large-model knowledge for anomaly analysis. Complementary latent-representation approaches include SteerVAD [6] and HiProbe [7], the latter probing frozen MLLM hidden states with a trained lightweight classifier. These developments motivate a distinction between language-mediated anomaly assessment and direct inference from latent representations. SphereVAD follows the latter direction, calibrating directional prototypes from synthetic references and performing geodesic inference on the unit hypersphere $\mathcal{S}^{D-1}$ without gradient-based training or autoregressive text generation.

Synthetic Data for Anomaly Calibration. Synthetic supervision has been explored in industrial anomaly detection, including simulated defects and image transformations [53, 24]. These methods use synthesized anomalies to train task-specific models. SphereVAD instead uses synthetic normal–abnormal references to calibrate a geometric decision rule over frozen MLLM features. Motivated by HiProbe [7], which demonstrates the utility of intermediate-layer representations with limited labeled examples, SphereVAD constructs directional prototypes from approximately 2,000 synthetic images without training an anomaly classifier. With the concentration parameter held fixed, vMF directional prototypes can be estimated through normalized cluster means. Synthetic samples therefore serve as feature-space calibration references rather than training data for representation learning.

3 Method

3.1 Overview and Problem Definition

Problem definition. Given a set of untrimmed test videos $\{V_1, \ldots, V_N\}$, each uniformly divided into T_i clips, the goal is to produce a per-frame anomaly score $s \in [0, 1]$ without any gradient-based training. The available resources comprise a pre-trained MLLM θ whose parameters are entirely frozen and a small synthetic calibration set $\mathcal{D}_{\text{syn}} = \{(x_j, y_j)\}_{j=1}^{M}$, $y_j \in \{\text{norm}, \text{abn}\}$, $M \approx 2000$.

Core idea. SphereVAD builds on a key observation: intermediate-layer features of MLLMs already encode rich anomaly semantics, yet the raw features are highly concentrated within a narrow cone (Figure 1(b)), severely under-utilizing the angular discriminative space. We recast anomaly detection as a vMF likelihood-ratio geodesic inference problem on the unit hypersphere $\mathcal{S}^{D-1}$, unleashing this latent discriminability through principled geometric reasoning.

As illustrated in Figure 2, the pipeline comprises four stages: **(a)** Feature preparation—extracting dual-stream intermediate-layer features and applying spherical centering to remove domain bias; **(b)** vMF center construction: unified Fréchet mean computation, logarithmic-map spherical centering, and spherical K-Means prototype calibration. **(c)** HSA—cross-video attention aggregation for feature

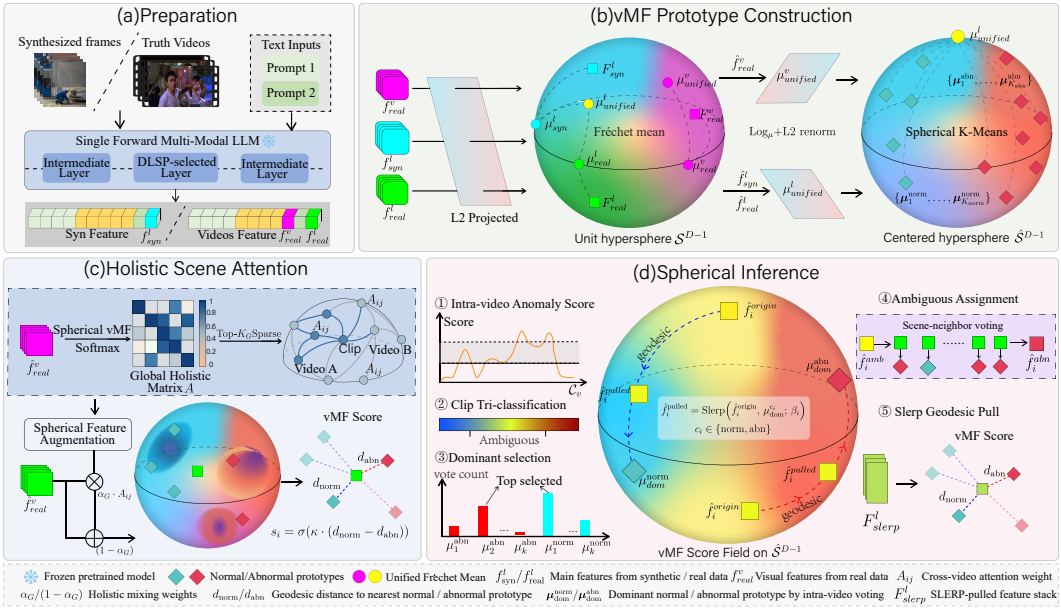


Figure 2: Overall pipeline of SphereVAD.

consistency; **(d)** vMF spherical inference—SGP on ambiguous clips followed by final scoring. The entire pipeline is training-free and involves only four core hyperparameters $(K_N, K_A, \alpha_G, \beta_{\text{base}})$.

Notation. Raw features are denoted f, ℓ_2-normalized features $\tilde{f}$, and spherically centered features $\hat{f}$. Superscripts l and v indicate the main feature (last token) and the visual feature (visual last), respectively.

3.2 Feature Extraction and Spherical Centering

Intermediate-layer feature extraction. From the frozen MLLM θ, we extract two complementary representations at the same DLSP-selected layer ℓ^*: the main feature $f^l \in \mathbb{R}^D$, the hidden state of the last sequence token, and the visual feature $f^v \in \mathbb{R}^D$, the hidden state of the final visual token. The layer-selection procedure is detailed in Appendix C.3. All features are projected onto the unit hypersphere by ℓ_2-normalization: $\tilde{f} = f/\|f\|_2 \in \mathcal{S}^{D-1}$.

Unified Fréchet mean and spherical centering. As discussed in Section 1, features from different data sources exhibit a systematic rotational bias on the hypersphere due to input-domain discrepancies (see Appendix F.3 for empirical evidence). Meanwhile, the raw features cluster within a narrow cone (Figure 1(b)), under-utilizing the angular discriminative capacity. Spherical centering addresses both issues at once.

We pool the main features from both domains to compute a unified Fréchet mean (details in Appendix A.1):

$$\boldsymbol{\mu}_{\text{unified}} = \text{FréchetMean}\big(\{\tilde{f}^l_{\text{syn}}\} \cup \{\tilde{f}^l_{\text{real}}\}\big). \tag{1}$$

Spherical centering is then performed with $\boldsymbol{\mu}_{\text{unified}}$ as the base point:

$$\hat{f} = \frac{\text{Log}_{\boldsymbol{\mu}_{\text{unified}}}(\tilde{f})}{\big\|\text{Log}_{\boldsymbol{\mu}_{\text{unified}}}(\tilde{f})\big\|_2} \in \mathcal{S}^{D-1}. \tag{2}$$

The logarithmic map [1] removes the mean-direction bias, and the subsequent ℓ_2-normalization projects the result back onto the sphere. This procedure is equivalent to spherical zero-centering, which unfolds the narrow-cone distribution over the full sphere and substantially expands the usable angular discrimination range. Visual features are independently centered with their own Fréchet mean $\boldsymbol{\mu}^v_{\text{unified}}$ (computed only from the test set, as detailed in Appendix C.4).

A natural question is whether this centering operation might inadvertently reduce the angular separation between normal and anomalous features. Proposition 1 shows that this is not the case.

Proposition 1 (Spherical centering does not reduce inter-class angular distance)**.** *Let $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \subset \mathcal{S}^{D-1}$, where each class follows* $\mathrm{vMF}(\boldsymbol{\mu}_c, \kappa_c)$*. Denote by ϕ the centering map of Eq.* (2)*. Under this idealized two-class vMF model, $d_{\mathrm{geo}}\big(\phi(\boldsymbol{\mu}_0), \phi(\boldsymbol{\mu}_1)\big) \geq d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$, with equality only in the degenerate case $d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) = \pi$. In practice, the bound holds approximately for finite samples when the empirical Fréchet mean is close to the population mean (Remark 2 in Appendix A.2).* Proof in Appendix A.2.

Furthermore, the choice of a unified Fréchet mean computed from both domains is not arbitrary. Proposition 2 guarantees that this mean naturally absorbs the rotational bias between domains, serving as an optimal shared reference point.

Proposition 2 (Unified Fréchet mean as rotational alignment)**.** *If the two domain distributions differ only by a rotation $R \in \mathrm{SO}(D)$ and share equal concentration and sample size, the unified Fréchet mean coincides with the geodesic midpoint of the domain-specific means. In the asymmetric case it still lies on the connecting geodesic, providing a reference that absorbs rotational bias from both sides.* Proof in Appendix A.3.

3.3 vMF Directional Prototype Construction and Scoring

With the centered synthetic calibration features in hand, we construct directional prototypes to serve as reference anchors on the hypersphere. Specifically, we perform spherical K-Means [14] clustering separately for the normal and anomalous classes, using cosine similarity for assignment and intra-cluster mean followed by ℓ_2-normalization for update. This yields K_N normal and K_A anomalous prototypes:

$$\{\boldsymbol{\mu}_k^{\mathrm{norm}}\}_{k=1}^{K_N}, \quad \{\boldsymbol{\mu}_k^{\mathrm{abn}}\}_{k=1}^{K_A} \subset \mathcal{S}^{D-1}. \tag{3}$$

These prototypes serve as the directional parameters of vMF distributions $\mathrm{vMF}(\mathbf{x}; \boldsymbol{\mu}, \kappa) \propto \exp(\kappa\, \boldsymbol{\mu}^\top \mathbf{x})$ [5]. Crucially, the synthetic data do not train a model but merely calibrate directional references on the sphere. Because vMF parameter estimation requires only the mean direction, the procedure is highly sample-efficient.

vMF likelihood-ratio scoring. For a clip feature $\hat{f}$, we compute the geodesic distance to the nearest prototype of each class, $d_c(\hat{f}) = \min_k \arccos(\hat{f}^\top \boldsymbol{\mu}_k^c)$, and define the anomaly score:

$$s = \sigma\big(\kappa\,(d_{\mathrm{norm}} - d_{\mathrm{abn}})\big) = \frac{\exp(-\kappa\, d_{\mathrm{abn}})}{\exp(-\kappa\, d_{\mathrm{norm}}) + \exp(-\kappa\, d_{\mathrm{abn}})}. \tag{4}$$

A clip geodesically closer to an anomalous prototype receives $s > 0.5$ and is deemed anomalous, while κ controls the decision sharpness.

3.4 Cross-Video Holistic Scene Attention

The prototypes constructed in Section 3.3 provide reliable scoring for clips whose features are well separated from the decision boundary. However, clips from visually similar scenes may receive inconsistent scores due to isolated noise. HSA addresses this by leveraging the visual features $\hat{f}^v$ to capture cross-video scene correlations and reinforce the consistency of the main features $\hat{f}^l$.

A sparse attention matrix $\mathbf{A} \in \mathbb{R}^{N_{\mathrm{total}} \times N_{\mathrm{total}}}$ is constructed from the pairwise cosine similarities of all test-clip visual features, with threshold sparsification, top-K_G truncation, and row-wise softmax normalization. The enhanced main feature is:

$$\hat{f}_i^{l,\mathrm{enh}} = (1 - \alpha_G)\, \hat{f}_i^l + \alpha_G \sum_j A_{ij}\, \hat{f}_j^l. \tag{5}$$

The result is ℓ_2-normalized back onto the sphere. This aggregation enables clips from visually similar scenes to reinforce each other across videos, suppressing isolated noise and promoting consistent anomaly detection. Initial vMF scores $\{s_i^{\mathrm{init}}\}$ are then obtained by applying Eq. (4) to the enhanced features.

3.5 vMF-Guided Spherical Geodesic Pulling

After initial scoring via Eq. (4), a subset of clips remains in the ambiguous zone near the decision boundary. SGP proposes a within-video, manifold-constrained feature correction mechanism that

leverages vMF prototype priors to adaptively adjust the feature directions of these ambiguous clips. The procedure is applied independently to each video and comprises the following steps (full algorithm and design rationale in Appendix D).

Adaptive tri-classification. Based on the initial scores $\{s_i^{\text{init}}\}$ obtained in Section 3.4, an ambiguity interval $[\rho_{\text{low}}, \rho_{\text{high}}]$ is adaptively determined so that clips within each video are partitioned into three groups: $\{\text{norm}, \text{amb}, \text{abn}\}$. The interval width adapts to the score distribution, narrowing when scores are clearly bimodal and widening when they concentrate near 0.5 (details in Appendix D.1).

Dominant prototype selection. Within each video, the clearly normal and clearly anomalous clips vote for their respective dominant prototypes from the prototype sets defined in Eq. (3). These dominant prototypes represent the prevailing normal and anomalous directions for the current video. If the number of anomalous clips falls below a minimum threshold, the video is treated as entirely normal (Appendix D.2).

Ambiguous-clip assignment. Rather than relying on its own unreliable score, each ambiguous clip is assigned a label via a sparse intra-video attention matrix $\mathbf{A}_H$ built from visual features. The attention-weighted aggregation of neighbouring clip features is compared with the dominant prototypes by cosine similarity, so that neighbour consensus resolves the ambiguity (Appendix D.3).

Score-adaptive geodesic pull. Once assigned, the feature of each ambiguous clip is moved along the geodesic toward its target prototype via Spherical Linear Interpolation (SLERP) [39]:

$$\hat{f}_i^{\text{pulled}} = \text{Slerp}\big(\hat{f}_i^{\,l,\text{enh}}, \boldsymbol{\mu}_{\text{target},i}^{c_i}, \beta_i\big). \tag{6}$$

Here, c_i is the assigned class, and $\boldsymbol{\mu}_{\text{target},i}^{c_i}$ is the corresponding dominant prototype selected for clip i. The pull strength $\beta_i = \beta_{\text{base}}(1 - \frac{1}{2}\hat{d}_i)$ depends on the normalized distance $\hat{d}_i \in [0, 1]$ from the decision boundary, defined in Eq. (43). A score closer to 0.5 ($\hat{d}_i \to 0$) yields the maximum pull β_{base}, whereas more confident scores receive a linearly decayed correction. SLERP traverses the great-circle arc at constant angular velocity, ensuring that the result remains strictly on $\mathcal{S}^{D-1}$ and thereby achieving manifold-constrained feature correction without distribution drift (Appendix D.4).

The corrected features are scored via Eq. (4) to produce the final clip-level anomaly scores, which are expanded to frame level and temporally smoothed with a Gaussian kernel [40].

4 Experiments

We validate SphereVAD on three large-scale video anomaly detection benchmarks, conducting SOTA comparison (§4.1), module ablation (§4.2), analysis of spherical geometry advantages (§4.3), backbone generality verification (§4.4), qualitative analysis (§4.5), and an extension to the weakly supervised setting (§4.6).

Datasets and Metrics. (1) **XD-Violence** [45] comprises 4,754 untrimmed videos covering 6 violent anomaly categories; we report frame-level AP (primary) and AUC. (2) **UCF-Crime** [40] consists of 1,900 surveillance videos spanning 13 crime categories; frame-level AUC is the standard metric. (3) **UBnormal** [2] is a synthetic open-set benchmark with pixel-level annotations across 29 anomaly types; we report frame-level AUC. (4) **ShanghaiTech** [30] contains 437 videos from 13 campus scenes. It is used only in the weakly supervised extension (§4.6), with the re-split of [55] (238 training and 199 test videos); we report frame-level AUC. Following common practice, all metrics are computed by concatenating frames from all test videos.

Implementation Details. We employ a frozen Qwen3.5 [41] as the default backbone. The feature extraction layer is selected as layer 31 via the procedure described in Appendix C.3. We synthesize approximately 2,000 calibration images ($\sim$1,000 normal/anomalous pairs) using only binary labels; no anomaly category information is leaked and there is no content overlap with any test set (details in Appendix B). All experiments in §4.1–§4.5 run on 4$\times$A100 GPUs at inference time with no gradient descent; the only exception is the weakly supervised extension in §4.6, where a lightweight head is trained on frozen features. Beyond offline batch processing, SphereVAD also supports an **online streaming mode** in which each incoming clip is scored immediately after feature extraction and spherical centering using a variant of M1 with synthetic-only centering (without HSA, SGP, or transductive real-data statistics). In this mode, the per-clip geometric inference is negligible ($<$0.1 ms), so throughput is dominated by the single MLLM forward pass, achieving **59.3 FPS** on a

Table 1: **VAD performance comparison.** ✓/✗ indicate zero-shot and training-free status. Best training-free results are in **bold**.

Method	Venue	Zero-shot	Train-free	UCF-Crime AUC(%)	UBnormal AUC(%)	XD-Violence AUC(%)	XD-Violence AP(%)
Sultani et al. [40]	CVPR'18	✗	✗	77.92	50.30	–	73.20
RTFM [43]	ICCV'21	✗	✗	83.31	64.94	–	77.81
UR-DMU [56]	AAAI'23	✗	✗	86.97	–	94.02	81.66
CLIP-TSA [21]	ICIP'23	✗	✗	87.58	–	–	82.19
MGFN [11]	AAAI'23	✗	✗	86.98	–	–	80.11
VadCLIP [48]	AAAI'24	✗	✗	88.02	–	–	84.51
OVVAD [46]	CVPR'24	✗	✗	86.40	62.94	–	66.53
STPrompt [47]	ACM MM'24	✗	✗	88.08	63.98	–	–
TPWNG [50]	CVPR'24	✗	✗	87.79	–	–	83.68
Holmes-VAU [54]	CVPR'25	✗	✗	88.96	56.77	–	87.68
π-VAD [32]	CVPR'25	✗	✗	**90.33**	–	–	85.37
RefineVAD [22]	AAAI'26	✗	✗	88.92	–	–	88.66
VERA [51]	CVPR'25	✗	✓	86.55	–	88.26	70.54
LAVIDA [12]	CVPR'26	✓	✗	82.18	**76.45**	–	**90.62**
CLIP [36]	ICML'21	✓	✓	53.16	–	38.21	17.83
LLaVA-1.5 [28]	CVPR'24	✓	✓	72.84	–	79.62	50.26
LAVAD [52]	CVPR'24	✓	✓	80.28	51.06	85.36	62.01
MCANet [13]	ICPR'24	✓	✓	82.47	62.94	87.43	69.72
GLM-4.1V-9B [15]	ArXiv'25	✓	✓	61.80	60.81	72.73	52.93
EventVAD [38]	ACM MM'25	✓	✓	82.03	–	87.51	64.04
Unified-VAD [27]	NeurIPS'25	✓	✓	84.28	69.02	91.34	68.07
VADTree [26]	NeurIPS'25	✓	✓	84.74	65.80	90.55	68.85
PANDA [49]	NeurIPS'25	✓	✓	84.89	75.78	–	70.16
SphereVAD (Online)	NeurIPS'26	✓	✓	78.31	71.62	92.48	81.37
SphereVAD (Offline)	NeurIPS'26	✓	✓	**86.38**	**76.46**	**95.74**	**86.99**

single A100 GPU (measured at ∼67.5 ms per 4-frame clip forward pass)—comfortably exceeding real-time requirements. A detailed computational cost analysis covering both deployment modes is provided in Appendix E. Full hyperparameter settings and sensitivity analyses are in Appendix G.

4.1 Comparison with State-of-the-Art Methods

As shown in Table 1, SphereVAD establishes state-of-the-art performance among zero-shot, training-free methods across all three benchmarks. On XD-Violence, it surpasses the previous best zero-shot training-free method (PANDA, AP 70.16%) by **+16.83 % AP** (and exceeds the weakly-supervised training-free method VERA, 70.54%, by **+16.45 % AP**), even outperforming established supervised methods such as VadCLIP (84.51%) and π-VAD (85.37%). While the task-specific fine-tuned zero-shot method LAVIDA achieves a higher AP (90.62%) on XD-Violence by updating model weights, SphereVAD remains strictly zero-parameter-update and tuning-free. On UCF-Crime, SphereVAD achieves **86.38% AUC**, advancing the zero-shot training-free ceiling (+1.49 % over PANDA, 84.89%) and remaining highly competitive with VERA (86.55%, which utilizes weak training-set labels despite being training-free), while narrowing the gap with the top supervised method (π-VAD, 90.33%) to within 4 points. On UBnormal, SphereVAD attains **76.46% AUC**, outperforming PANDA (75.78%) and matching the fine-tuned LAVIDA (76.45%) without any task-specific gradient descent, while substantially surpassing supervised baselines (e.g., RTFM 64.94%, STPrompt 63.98%). The XD-Violence gains are particularly striking because this dataset contains diverse violence types that benefit from the cross-video HSA mechanism.

4.2 Ablation Studies

Table 2 presents the incremental contribution of each module across all three benchmarks. Spherical centering accounts for the largest single gain on UCF-Crime and UBnormal. On UCF-Crime it raises AUC from 75.42% to 81.20% (+5.78 pp), on UBnormal from 71.13% to 75.49% (+4.36 pp), and on XD-Violence from 80.19% to 82.03% (+1.84 % AP), confirming the necessity of unfolding narrow-cone features onto the full sphere. Adding HSA (M2) brings +2.65 % AP on XD-Violence (representing the largest single gain on this dataset) and +0.48 % AUC on UBnormal, confirming the

Table 2: **Component-wise ablation.** Each row adds one module to its predecessor. "Eucl." denotes Euclidean distance-ratio scoring; "vMF" denotes vMF log-likelihood-ratio scoring. Δ is the improvement of vMF over Euclidean scoring at each stage. Primary metrics are underlined. UBnormal does not use SGP, so M2 and M3 share the same results.

	Configuration	XD-Violence AP (%)			UCF-Crime AUC (%)			UBnormal AUC (%)		
		Eucl.	vMF	Δ	Eucl.	vMF	Δ	Eucl.	vMF	Δ
M0a	Raw features	80.19	–	–	75.42	–	–	71.13	–	–
M0b	+ Spherical center	82.03	–	–	81.20	–	–	75.49	–	–
M1	+ vMF scoring	82.03	82.51	+0.48	81.20	81.68	+0.48	75.49	75.98	+0.49
M2	+ HSA	85.01	85.16	+0.15	82.71	83.06	+0.35	76.01	76.46	+0.45
M3	+ SGP	85.60	**86.99**	+1.39	84.80	**86.38**	+1.58	76.01†	**76.46**†	+0.45

† UBnormal videos average $\sim$12 seconds, too short for intra-video SGP; M3 results are identical to M2.

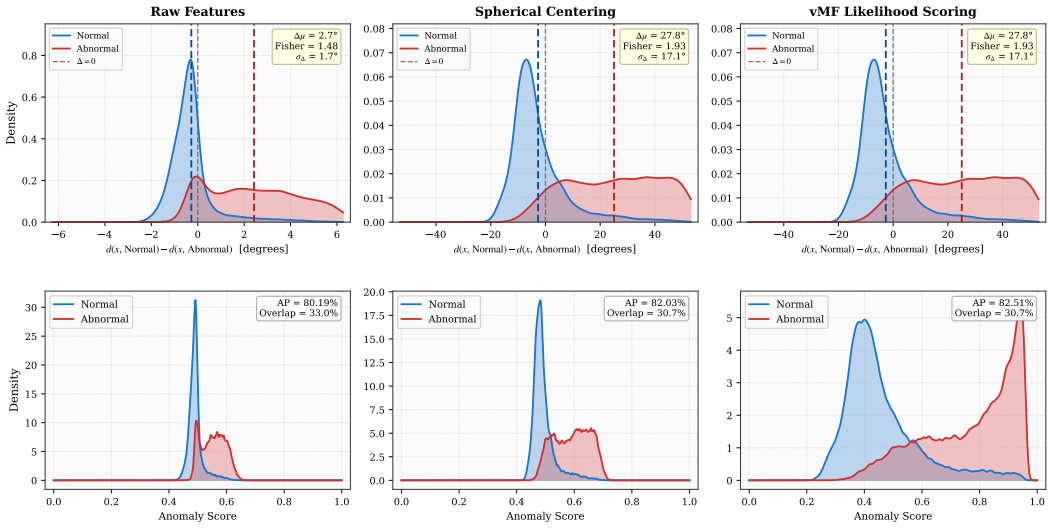


Figure 3: Progressive improvement of feature discriminative structure. **Top:** geodesic distance difference $\Delta = d(\mathbf{x}, \text{Normal}) - d(\mathbf{x}, \text{Abnormal})$. **Bottom:** anomaly score distributions. Columns from left to right: Raw / Spherical Centering / vMF Scoring.

value of cross-video scene-consistency priors. SGP (M3) further yields $+3.32\,\%$ AUC on UCF-Crime and $+1.83\,\%$ AP on XD-Violence, demonstrating that directional correction of ambiguous transitional segments is critical for long surveillance videos. Note that UBnormal does not employ SGP because its test videos average only $\sim$12 seconds each, providing insufficient temporal context for the intra-video tri-classification and neighbour-consensus mechanisms that SGP relies on. Consequently M2 and M3 share identical results. Cumulatively, M1$\rightarrow$M3 improves by $+4.48\,\%$ AP (XD-Violence) and $+4.70\,\%$ AUC (UCF-Crime).

4.3 Advantages of Spherical Geometry

4.3.1 vMF Scoring vs. Distance-Ratio Scoring

Across all three benchmarks and all pipeline stages, vMF log-likelihood-ratio scoring consistently outperforms Euclidean distance-ratio scoring (Table 2, Δ columns). In stage M1, the calibrated concentration κ acts as an optimal temperature parameter that bounds anomalous extreme values prior to temporal Gaussian smoothing, producing a consistent $+0.48\% \sim +0.49\%$ gain over raw distance ratios. The advantage is amplified when moving to stage M3 ($+1.39\,\%$ AP on XD-Violence and $+1.58\,\%$ AUC on UCF-Crime), demonstrating that directional geodesic pulling and the bounded probability metric synergistically refine clips near the decision boundary.

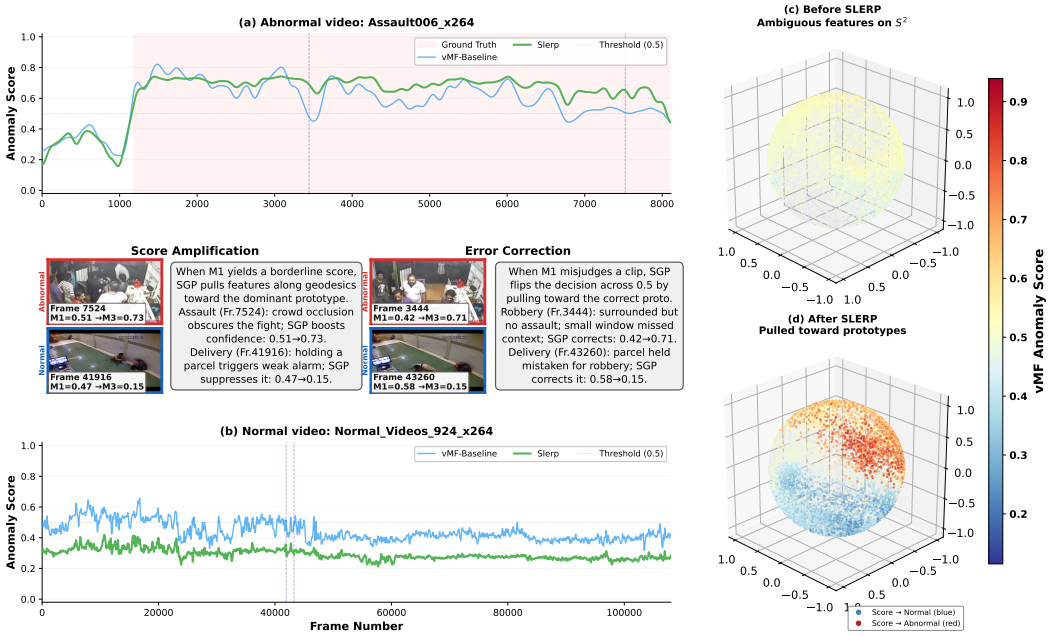


Figure 4: Qualitative results of SGP. **Left (a)(b):** Frame-level score curves (vMF-Baseline / SGP vs. GT) for two videos. **Right (c)(d):** Spherical visualization before vs. after pulling.

Table 3: **Backbone generality.** SphereVAD with four different MLLM backbones. The first row uses the same Qwen3.5 backbone but directly prompts its language head to rate anomaly severity, without our projection framework.

Backbone / Method	XD-Violence AP(%)	UCF-Crime AUC(%)	UBnormal AUC(%)
Qwen3.5 direct inference†	58.70	72.64	55.74
Llava-ov-1.5 [3]	76.93	79.36	69.83
InternVL3 [57]	80.41	84.76	72.39
Qwen3-VL [4]	82.81	83.42	74.51
Qwen3.5 [41]	**86.99** (+28.29)	**86.38** (+13.74)	**76.46** (+20.72)

† Language-only baseline: prompting Qwen3.5's language head to score anomaly severity without SphereVAD.

4.3.2 Visualization of Feature Discriminative Structure

Figure 3 visualizes how each stage reshapes the feature discriminative structure. Raw features exhibit near-complete overlap between normal and anomalous distance-difference distributions ($\Delta\mu$=2.7°, σ_Δ=1.7°, Fisher=1.48, score overlap 33.0%, AP=80.19%). Spherical centering expands $\Delta\mu$ to 27.8° (σ_Δ=17.1°, Fisher=1.93) and reduces score overlap to 30.7% (AP=82.03%), confirming its role as the impactful component. It unfolds features compressed within a narrow cone onto the full sphere, recovering wasted angular discriminative capacity. vMF scoring preserves the same geodesic separation ($\Delta\mu$=27.8°, Fisher=1.93) while nonlinearly reshaping the score distribution through the vMF log-likelihood ratio, lifting AP to 82.51%, consistent with the ablation results in Table 2.

4.4 Backbone Generality

As shown in Table 3, SphereVAD produces effective results across all four MLLM backbones, demonstrating that the spherical geometric inference framework is backbone-agnostic. Even the weakest backbone (LLaVA-OV-1.5) achieves 76.93% AP on XD, 79.36% AUC on UCF, and 69.83% on UBnormal, while the best backbone (Qwen3.5) reaches 86.99%, 86.38%, and 76.46%, respectively. Comparing with the LLM direct inference baseline (58.70% / 72.64% / 55.74%), SphereVAD with the same Qwen3.5 backbone yields improvements of +28.29%, +13.74%, and +20.72% on the three benchmarks. This reveals the bottleneck of the language decoding head in fine-grained anomaly understanding—what matters is not what the model "says", but what the model "sees".

Table 4: **Frame-level AUC (%) on ShanghaiTech.** Upper block: trained with video-level labels. Lower block: no training. Best in each block is in **bold**.

Method	Venue	Supervision	AUC(%)
Sultani et al. [40]	CVPR'18	WS (C3D)	86.30
RTFM [43]	ICCV'21	WS (I3D)	97.21
S3R [44]	ECCV'22	WS (I3D)	97.48
CLIP-TSA [21]	ICIP'23	WS (CLIP)	97.74
VadCLIP [48]	AAAI'24	WS (CLIP)	**97.80**
SphereVAD + MIL (ours)	NeurIPS'26	WS (frozen Qwen3.5)	96.38
LAVAD [52]	CVPR'24	Training-free	82.10
MCANet [13]	ICPR'24	Training-free	83.40
SphereVAD (offline)	NeurIPS'26	Training-free	**86.07**

4.5 Qualitative Analysis

Figure 4 illustrates temporal and geometric effects of SGP. For the anomalous video (a), the vMF-Baseline (M1) captures anomalous intervals but fluctuates; SGP (M3) corrects boundary segments, aligning the curve with the ground truth (e.g., Frame 7524: $0.51 \rightarrow 0.73$, and Frame 3444: $0.42 \rightarrow 0.71$). For the normal video (b), SGP suppresses M1's false positives (e.g., Frame 41916: $0.47 \rightarrow 0.15$ and Frame 43260: $0.58 \rightarrow 0.15$) by pulling ambiguous representations toward the dominant normal prototype. The spherical views (c–d) confirm the mechanism: before pulling, ambiguous features scatter between prototypes with mixed labels; after pulling, they migrate along geodesics toward assigned prototypes, yielding clearly separated normal (blue) and anomalous (red) clusters.

4.6 Extension to the Weakly Supervised Setting

We further investigate whether spherical centering benefits downstream learning when video-level labels are available. On ShanghaiTech [30], using the re-split of [55] with 238 training and 199 test videos, we freeze Qwen3.5 and train only a lightweight temporal MIL head [40, 43] on the centered features. The unified Fr'echet mean is computed from synthetic calibration features and ShanghaiTech training features, without using test features; full implementation details are provided in Appendix G.5. As shown in Table 4, SphereVAD+MIL achieves 96.38% frame-level AUC, improving over its training-free counterpart by 10.31 percentage points and falling within 1.42 percentage points of VadCLIP (97.80%). Training the same head on features with only ℓ_2-normalization yields 92.85% AUC, demonstrating a 3.53-percentage-point gain from spherical centering. This extension involves supervised training and is reported separately from the training-free results in Table 1; it does not support the paper's training-free claims.

5 Conclusion

We propose SphereVAD, a training-free video anomaly detection framework that recasts anomaly discrimination as vMF likelihood-ratio geodesic inference on the unit hypersphere. SphereVAD comprises three complementary components: unified Fréchet mean spherical centering, cross-video holistic scene attention, and manifold-constrained spherical geodesic pulling, all operating without gradient descent. Using only $\sim$2,000 synthetic calibration images, SphereVAD achieves state-of-the-art results among zero-shot, training-free approaches on XD-Violence, UCF-Crime, and UBnormal, while remaining competitive with weakly supervised methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 62301621, 62672529, 62301323, 62625602, and 62441619), Shenzhen Science and Technology Program (Nos. 20231121172359002, 2023A008, SYSRD20250529113401002, and KQTD20221101093559018), Shenzhen General Research Project (No. JCYJ20241202125904007), and Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398).

Competing interests. The authors declare no competing interests.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

[2] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20143–20153, 2022.

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005.

[6] Zhaolin Cai, Fan Li, Huiyu Duan, Lijun He, and Guangtao Zhai. Steering and rectifying latent representation manifolds in frozen multi-modal llms for video anomaly detection. *arXiv preprint arXiv:2602.24021*, 2026.

[7] Zhaolin Cai, Fan Li, Ziwei Zheng, and Yanjun Qin. Hiprobe-vad: Video anomaly detection via hidden states probing in tuning-free multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 592–601, 2025.

[8] 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*, pages 18319–18329, 2024.

[9] Peng Chen, Chao Huang, Yunkang Cao, Chengliang Liu, Wei Wang, Wenqiang Wang, Mingbo Yang, Li Shen, Wenqi Ren, and Xiaochun Cao. Towards explainable industrial anomaly detection via knowledge-guided latent reasoning. *arXiv preprint arXiv:2602.09850*, 2026.

[10] Peng Chen, Fangjun Huang, and Chao Huang. Dyc-clip: Dynamic context-aware multi-modal prompt learning for zero-shot anomaly detection. *Pattern Recognition*, page 113215, 2026.

[11] 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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 387–395, 2023.

[12] Zunkai Dai, Ke Li, Jiajia Liu, Jie Yang, and Yuanyuan Qiao. No need for real anomaly: Mllm empowered zero-shot video anomaly detection. *arXiv preprint arXiv:2602.19248*, 2026.

[13] Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. Mcanet: Multimodal caption aware training-free video anomaly detection via large language model. In *International Conference on Pattern Recognition*, pages 362–379. Springer, 2024.

[14] Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175, 2001.

[15] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

[16] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. Weakly supervised video anomaly detection via self-guided temporal discriminative transformer. *IEEE Transactions on Cybernetics*, 54(5):3197–3210, 2022.

[17] 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 *Forty-second International Conference on Machine Learning*, 2025.

[18] Chao Huang, Benfeng Wang, Wei Wang, Jie Wen, Chengliang Liu, Li Shen, and Xiaochun Cao. Vad-r1: Towards video anomaly reasoning via perception-to-cognition chain-of-thought. *Advances in neural information processing systems*, 38:118486–118518, 2026.

[19] Chao Huang, Jie Wen, Chengliang Liu, and Yabo Liu. Long short-term dynamic prototype alignment learning for video anomaly detection. In *IJCAI*, pages 866–874, 2024.

[20] Chao Huang, Jie Wen, Yong Xu, Qiuping Jiang, Jian Yang, Yaowei Wang, and David Zhang. Self-supervised attentive generative adversarial networks for video anomaly detection. *IEEE transactions on neural networks and learning systems*, 34(11):9389–9403, 2022.

[21] Hyekang Kevin Joo, Khoa Vo, Kashu Yamazaki, and Ngan Le. Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 3230–3234. IEEE, 2023.

[22] Junhee Lee, ChaeBeen Bang, MyoungChul Kim, and MyeongAh Cho. Refinevad: Semantic-guided feature recalibration for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5899–5907, 2026.

[23] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.

[24] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023.

[26] Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *arXiv preprint arXiv:2510.22693*, 2025.

[27] Dongheng Lin, Mengxue Qu, Kunyang Han, Jianbo Jiao, Xiaojie Jin, and Yunchao Wei. A unified reasoning framework for holistic zero-shot video anomaly analysis. *arXiv preprint arXiv:2511.00962*, 2025.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[29] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *European conference on computer vision*, pages 125–141. Springer, 2020.

[30] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *2017 IEEE international conference on computer vision (ICCV)*, pages 341–349. IEEE, 2017.

[31] Junxiao Ma, Jingjing Wang, Jiamin Luo, Peiying Yu, and Guodong Zhou. Sherlock: Towards multi-scene video abnormal event extraction and localization via a global-local spatial-sensitive llm. In *Proceedings of the ACM on Web Conference 2025*, pages 4004–4013, 2025.

[32] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and Francois Bremond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24265–24274, 2025.

[33] Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

[34] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12173–12182, 2020.

[35] Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

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

[37] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2293–2312, 2020.

[38] Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2586–2595, 2025.

[39] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985.

[40] 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*, pages 6479–6488, 2018.

[41] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026.

[42] Kamalakar Vijay Thakare, Yash Raghuwanshi, Debi Prosad Dogra, Heeseung Choi, and Ig-Jae Kim. Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 5541–5550, 2023.

[43] 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*, pages 4975–4986, 2021.

[44] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022.

[45] 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 *European conference on computer vision*, pages 322–339. Springer, 2020.

[46] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024.

[47] Peng Wu, Xuerong Zhou, Guansong Pang, Zhiwei Yang, Qingsen Yan, Peng Wang, and Yanning Zhang. Weakly supervised video anomaly detection and localization with spatio-temporal prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9301–9310, 2024.

[48] 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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 6074–6082, 2024.

[49] Zhiwei Yang, Chen Gao, and Mike Zheng Shou. Panda: Towards generalist video anomaly detection via agentic ai engineer. *arXiv preprint arXiv:2509.26386*, 2025.

[50] 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*, pages 18899–18908, 2024.

[51] Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8679–8688, 2025.

[52] Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2024.

[53] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021.

[54] Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the computer vision and pattern recognition conference*, pages 13843–13853, 2025.

[55] 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 *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 1237–1246. IEEE, 2019.

[56] Hang Zhou, Junqing Yu, and Wei Yang. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3769–3777, 2023.

[57] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

The appendix is organised as follows. Appendix A provides theoretical proofs. Appendix B describes the synthetic calibration dataset. Appendix C details prompt design and feature extraction. Appendix D presents the full SGP algorithm. Appendix E reports computational cost analysis. Appendix F shows additional visualizations. Appendix G contains hyperparameter settings and sensitivity analyses. Appendix H provides the reproducibility statement and symbol table. Appendix I discusses limitations. Appendix J addresses broader impact.

A Theoretical Analysis

This section provides the complete mathematical machinery underlying the spherical centering framework of SphereVAD. Section A.1 details the Karcher iteration used to compute the unified Fréchet mean (Eq. (1)), including convergence guarantees and step-size justification. Section A.2 proves Proposition 1 (spherical centering expands inter-class angular distance). Section A.3 proves Proposition 2 (the unified Fréchet mean provides symmetric rotational alignment).

A.1 Fréchet Mean via Karcher Iteration

The Fréchet mean on a Riemannian manifold $\mathcal{M}$ generalises the Euclidean centre of mass. For a set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}^{D-1}$, it is defined as

$$\boldsymbol{\mu}^* = \arg\min_{\boldsymbol{\mu} \in \mathcal{S}^{D-1}} \sum_{i=1}^{N} d_{\text{geo}}^2(\boldsymbol{\mu}, \mathbf{x}_i), \tag{7}$$

where $d_{\text{geo}}(\boldsymbol{\mu}, \mathbf{x}) = \arccos(\boldsymbol{\mu}^\top \mathbf{x})$ is the geodesic (great-circle) distance on the unit hypersphere.

A.1.1 Logarithmic and Exponential Maps

The logarithmic and exponential maps provide the bridge between the sphere and its tangent space, enabling Riemannian optimisation.

Logarithmic map. For a base point $\boldsymbol{\mu} \in \mathcal{S}^{D-1}$ and a target point $\mathbf{x} \in \mathcal{S}^{D-1}$ with $\mathbf{x} \neq \pm\boldsymbol{\mu}$, the logarithmic map yields the unique tangent vector $\mathbf{v} \in T_{\boldsymbol{\mu}}\mathcal{S}^{D-1}$ such that the geodesic starting at $\boldsymbol{\mu}$ with initial velocity $\mathbf{v}$ reaches $\mathbf{x}$ at unit time:

$$\text{Log}_{\boldsymbol{\mu}}(\mathbf{x}) = \frac{\theta}{\sin\theta}\big(\mathbf{x} - \cos\theta\, \boldsymbol{\mu}\big), \qquad \theta = \arccos\big(\boldsymbol{\mu}^\top \mathbf{x}\big). \tag{8}$$

Crucially, $\|\text{Log}_{\boldsymbol{\mu}}(\mathbf{x})\| = \theta = d_{\text{geo}}(\boldsymbol{\mu}, \mathbf{x})$ and $\text{Log}_{\boldsymbol{\mu}}(\mathbf{x})^\top \boldsymbol{\mu} = 0$ (the result is orthogonal to $\boldsymbol{\mu}$, confirming it lies in $T_{\boldsymbol{\mu}}\mathcal{S}^{D-1}$). When $\theta \to 0$, we recover $\text{Log}_{\boldsymbol{\mu}}(\mathbf{x}) \to \mathbf{x} - \boldsymbol{\mu}$ (the Euclidean difference).

Exponential map. For a tangent vector $\mathbf{v} \in T_{\boldsymbol{\mu}}\mathcal{S}^{D-1}$ (i.e., $\mathbf{v}^\top \boldsymbol{\mu} = 0$), the exponential map traces the geodesic from $\boldsymbol{\mu}$ in direction $\mathbf{v}$ for unit time:

$$\text{Exp}_{\boldsymbol{\mu}}(\mathbf{v}) = \cos\big(\|\mathbf{v}\|\big)\, \boldsymbol{\mu} + \sin\big(\|\mathbf{v}\|\big)\, \frac{\mathbf{v}}{\|\mathbf{v}\|}. \tag{9}$$

One can verify that $\|\text{Exp}_{\boldsymbol{\mu}}(\mathbf{v})\| = 1$ and $d_{\text{geo}}(\boldsymbol{\mu}, \text{Exp}_{\boldsymbol{\mu}}(\mathbf{v})) = \|\mathbf{v}\|$, confirming that the exponential map preserves geodesic distance in the radial direction. When $\|\mathbf{v}\| \to 0$, we recover $\text{Exp}_{\boldsymbol{\mu}}(\mathbf{v}) \to \boldsymbol{\mu} + \mathbf{v}$.

A.1.2 Karcher Iteration Algorithm

The Karcher iteration (also called Riemannian gradient descent for the Fréchet objective) proceeds as follows:

1. **Initialisation.** Compute the Euclidean mean and project onto the sphere: $\boldsymbol{\mu}^{(0)} = \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_i \mathbf{x}_i$. This provides a consistent starting point in the vicinity of the true Fréchet mean.

2. **Gradient computation and update.** At iteration t, compute the Riemannian gradient of the Fréchet objective:

$$\mathbf{g}^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Log}_{\boldsymbol{\mu}^{(t)}}(\mathbf{x}_i). \tag{10}$$

This is the (negative) Riemannian gradient of $F(\boldsymbol{\mu}) = \sum_i d^2_{\mathrm{geo}}(\boldsymbol{\mu}, \mathbf{x}_i)$ divided by $2N$; it points from $\boldsymbol{\mu}^{(t)}$ toward the tangent-space centre of mass. Update the estimate via the exponential map:

$$\boldsymbol{\mu}^{(t+1)} = \mathrm{Exp}_{\boldsymbol{\mu}^{(t)}}\big(\eta\, \mathbf{g}^{(t)}\big), \tag{11}$$

where $\eta > 0$ is the step size.

3. **Convergence check.** Terminate when $\|\mathbf{g}^{(t)}\| < \epsilon$ or a maximum number of iterations is reached. At convergence, $\mathbf{g}^* = \mathbf{0}$, which is the necessary and sufficient first-order condition for the Fréchet mean.

A.1.3 Convergence Guarantee and Step-Size Justification

Theorem 1 (Convergence of Karcher iteration[35])**.** *Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}^{D-1}$ be contained within an open geodesic ball $B(\mathbf{p}, r)$ with radius $r < \pi/2$. Then:*

(i) (Existence and uniqueness.) *The Fréchet mean $\boldsymbol{\mu}^*$ exists, is unique, and lies within $B(\mathbf{p}, r)$.*

(ii) (Linear convergence.) *The Karcher iteration with step size $\eta = 1$ converges to $\boldsymbol{\mu}^*$ at a linear rate: $d_{\mathrm{geo}}(\boldsymbol{\mu}^{(t+1)}, \boldsymbol{\mu}^*) \le c\, d_{\mathrm{geo}}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\mu}^*)$ for some $c < 1$ depending on the curvature and data spread.*

Step size $\eta = 1$. On the unit sphere with constant positive curvature $\kappa_{\mathrm{sec}} = 1$, the squared geodesic distance $d^2_{\mathrm{geo}}(\boldsymbol{\mu}, \mathbf{x})$ is geodesically convex for $d_{\mathrm{geo}}(\boldsymbol{\mu}, \mathbf{x}) < \pi/2$ [1]. Hence the Fréchet objective $F(\boldsymbol{\mu})$ is a sum of geodesically convex functions, and a full Riemannian gradient step ($\eta = 1$) is guaranteed to decrease F at each iteration. Unlike Euclidean gradient descent, overshooting cannot occur because the exponential map automatically constrains the update to lie on $\mathcal{S}^{D-1}$.

Practical convergence in SphereVAD. In our setting, MLLM intermediate-layer features before centering have angular standard deviation $\approx 1.7^\circ$, so they are contained within a geodesic ball of radius $r \lesssim 15^\circ \ll 90^\circ$. The convergence condition $r < \pi/2$ is thus satisfied by a wide margin. We use $T_{\max} = 5$ iterations and $\epsilon = 10^{-7}$; empirically, the gradient norm $\|\mathbf{g}^{(t)}\|$ drops below ϵ within 3 iterations.

A.2 Proof of Proposition 1: Spherical Centering Expands Inter-Class Angular Distance

For convenience, we restate the proposition:

Proposition (Proposition 1, restated)**.** *Let $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \subset \mathcal{S}^{D-1}$, where class $c \in \{0, 1\}$ follows $\mathrm{vMF}(\boldsymbol{\mu}_c, \kappa_c)$ with mixture weight $\pi_c > 0$. Denote the population Fréchet mean of the mixture by $\boldsymbol{\mu}$, and define the centering map $\phi(\mathbf{x}) = \mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x}) / \big\|\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})\big\|$. Then*

$$d_{\mathrm{geo}}\big(\phi(\boldsymbol{\mu}_0),\, \phi(\boldsymbol{\mu}_1)\big) \;\ge\; d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1). \tag{12}$$

Moreover, the inequality is strict whenever $0 < d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) < \pi$.

The proof proceeds in three steps. We first establish a key geometric lemma (Lemma 1), then combine it with the anti-parallel property of the logarithmic map to obtain the result.

Lemma 1 (Fréchet mean of a vMF mixture lies on the inter-class geodesic)**.** *Under the conditions of Proposition 1, the population Fréchet mean $\boldsymbol{\mu}$ lies on the minimal geodesic segment from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_1$ in $\mathcal{S}^{D-1}$. That is, there exists $t_0 \in (0, 1)$ such that*

$$\boldsymbol{\mu} = \mathrm{Slerp}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, t_0) = \frac{\sin\big((1 - t_0)\alpha\big)}{\sin \alpha}\, \boldsymbol{\mu}_0 + \frac{\sin(t_0 \alpha)}{\sin \alpha}\, \boldsymbol{\mu}_1, \tag{13}$$

where $\alpha = d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) \in (0, \pi)$.

Proof of Lemma 1. Define the two-dimensional subspace $V = \operatorname{span}\{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1\} \subset \mathbb{R}^D$ and its orthogonal complement $V^\perp$ (dimension $D-2 \geq 1$ for $D \geq 3$; the $D = 2$ case is treated separately below).

Consider any orthogonal transformation $Q \in \mathrm{O}(D)$ that fixes V pointwise and acts arbitrarily on $V^\perp$:

$$Q\mathbf{v} = \mathbf{v} \;\; \forall \mathbf{v} \in V, \qquad Q\big|_{V^\perp} \in \mathrm{O}(D-2). \tag{14}$$

Since $Q\boldsymbol{\mu}_c = \boldsymbol{\mu}_c$ for $c \in \{0,1\}$, and the vMF density $\mathrm{vMF}(\mathbf{x}; \boldsymbol{\mu}_c, \kappa_c) \propto \exp(\kappa_c \boldsymbol{\mu}_c^\top \mathbf{x})$ depends on $\mathbf{x}$ only through $\boldsymbol{\mu}_c^\top \mathbf{x}$, we have $\boldsymbol{\mu}_c^\top (Q\mathbf{x}) = (Q^\top \boldsymbol{\mu}_c)^\top \mathbf{x} = \boldsymbol{\mu}_c^\top \mathbf{x}$. Thus the distribution of $Q\mathbf{x}$ under $\mathrm{vMF}(\boldsymbol{\mu}_c, \kappa_c)$ equals the distribution of $\mathbf{x}$; in other words, each class distribution is Q-invariant.

The Fréchet objective of the mixture is

$$F(\boldsymbol{\mu}) = \sum_{c \in \{0,1\}} \pi_c \, \mathbb{E}_{\mathbf{x} \sim \mathrm{vMF}(\boldsymbol{\mu}_c, \kappa_c)} \big[d_{\mathrm{geo}}^2(\boldsymbol{\mu}, \mathbf{x}) \big].$$

The Q-invariance of the class distributions gives $F(Q\boldsymbol{\mu}) = F(\boldsymbol{\mu})$ for every Q of the form (14). Since F is strictly geodesically convex on any open hemisphere (Theorem 1), the minimiser $\boldsymbol{\mu}$ is unique, hence $Q\boldsymbol{\mu} = \boldsymbol{\mu}$ for all such Q. The fixed-point set of all such Q is exactly $V \cap \mathcal{S}^{D-1}$—the great circle through $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$. Therefore, $\boldsymbol{\mu}$ lies on this great circle. Because $\boldsymbol{\mu}$ minimises rather than maximises F, it lies on the *shorter* arc (the minimal geodesic segment) between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, yielding (13) for some $t_0 \in (0,1)$.

Case $D = 2$. When $D = 2$, the subspace $V = \mathbb{R}^2$ and $V^\perp = \{\mathbf{0}\}$, so the symmetry argument produces only the trivial transformation $Q = I$. Nevertheless, on $\mathcal{S}^1$ the Fréchet objective restricted to the great circle is a sum of $\cos^{-2}$-type functions, which is strictly convex on any arc shorter than π. The minimiser is the weighted circular Fréchet mean, which lies on the shorter arc between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ whenever $\alpha = d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) < \pi$ [35]. □

Proof of Proposition 1. We proceed in three steps.

Step 1: $\boldsymbol{\mu}$ lies on the inter-class geodesic. By Lemma 1, $\boldsymbol{\mu} = \gamma(t_0)$ for some $t_0 \in (0,1)$, where $\gamma : [0,1] \to \mathcal{S}^{D-1}$ is the unit-speed geodesic from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_1$. Let $\theta_0 = d_{\mathrm{geo}}(\boldsymbol{\mu}, \boldsymbol{\mu}_0) = t_0 \alpha > 0$ and $\theta_1 = d_{\mathrm{geo}}(\boldsymbol{\mu}, \boldsymbol{\mu}_1) = (1-t_0)\alpha > 0$, so that $\theta_0 + \theta_1 = \alpha$.

Step 2: The logarithmic maps at $\boldsymbol{\mu}$ are anti-parallel. Let $\mathbf{u} \in T_{\boldsymbol{\mu}} \mathcal{S}^{D-1}$ be the unit tangent vector to γ at $\boldsymbol{\mu}$ pointing from $\boldsymbol{\mu}_0$ toward $\boldsymbol{\mu}_1$. Since $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ lie on the same great circle through $\boldsymbol{\mu}$ but on *opposite sides* of $\boldsymbol{\mu}$, the logarithmic maps satisfy:

$$\mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0) = -\theta_0 \, \mathbf{u}, \qquad \mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_1) = +\theta_1 \, \mathbf{u}. \tag{15}$$

These two tangent vectors point in exactly opposite directions in $T_{\boldsymbol{\mu}} \mathcal{S}^{D-1}$. This follows directly from the definition of the logarithmic map: $\mathrm{Log}_{\boldsymbol{\mu}}(\mathbf{x})$ is the initial velocity of the unique geodesic from $\boldsymbol{\mu}$ to $\mathbf{x}$, scaled to have norm $d_{\mathrm{geo}}(\boldsymbol{\mu}, \mathbf{x})$. Since $\boldsymbol{\mu}_0$ is reached by travelling along γ in the backward direction ($-\mathbf{u}$) and $\boldsymbol{\mu}_1$ by travelling in the forward direction ($+\mathbf{u}$), Eq. (15) follows.

Step 3: After ℓ_2-normalisation, the centred class means are antipodal. Applying the centering map ϕ:

$$\phi(\boldsymbol{\mu}_0) = \frac{\mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0)}{\|\mathrm{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_0)\|} = \frac{-\theta_0 \, \mathbf{u}}{\theta_0} = -\mathbf{u}, \qquad \phi(\boldsymbol{\mu}_1) = \frac{+\theta_1 \, \mathbf{u}}{\theta_1} = +\mathbf{u}. \tag{16}$$

Both $\phi(\boldsymbol{\mu}_0) = -\mathbf{u}$ and $\phi(\boldsymbol{\mu}_1) = +\mathbf{u}$ are unit vectors in $\mathbb{R}^D$ (hence on $\mathcal{S}^{D-1}$), and they are antipodal:

$$d_{\mathrm{geo}}\big(\phi(\boldsymbol{\mu}_0), \, \phi(\boldsymbol{\mu}_1)\big) = \arccos\big((-\mathbf{u})^\top \mathbf{u}\big) = \arccos(-1) = \pi. \tag{17}$$

Since $d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) = \alpha \leq \pi$, we conclude

$$d_{\mathrm{geo}}\big(\phi(\boldsymbol{\mu}_0), \, \phi(\boldsymbol{\mu}_1)\big) = \pi \;\geq\; \alpha = d_{\mathrm{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1),$$

and the inequality is strict whenever $\alpha < \pi$. □

Remark 1 (Geometric intuition)**.** *The geometric picture can be summarised as follows. The Fréchet mean $\boldsymbol{\mu}$ lies on the geodesic between the two class means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, separated by arc length α. The logarithmic maps at $\boldsymbol{\mu}$ produce anti-parallel tangent vectors $-\theta_0\mathbf{u}$ and $+\theta_1\mathbf{u}$. After ℓ_2-normalisation, the centred class means $\phi(\boldsymbol{\mu}_0) = -\mathbf{u}$ and $\phi(\boldsymbol{\mu}_1) = +\mathbf{u}$ become* antipodal*, achieving maximal geodesic separation $\pi \geq \alpha$. The centering map discards the radial (distance) information and retains only the directional information relative to $\boldsymbol{\mu}$, thereby amplifying the angular contrast between the two classes.*

Remark 2 (Finite-sample case)**.** *In practice, the centering map uses the* empirical *unified Fréchet mean $\hat{\boldsymbol{\mu}}$ computed from N samples rather than the population mean $\boldsymbol{\mu}$. Standard concentration results on the sphere [35] guarantee that $d_{\text{geo}}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = O_p(N^{-1/2})$. As $\hat{\boldsymbol{\mu}} \to \boldsymbol{\mu}$, the angle A at $\hat{\boldsymbol{\mu}}$ in the geodesic triangle $(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$ converges to π by continuity of the logarithmic map. More precisely, using the spherical law of cosines:*

$$\cos A = \frac{\cos\alpha - \cos\hat{\theta}_0 \cos\hat{\theta}_1}{\sin\hat{\theta}_0 \sin\hat{\theta}_1}, \tag{18}$$

where $\hat{\theta}_c = d_{\text{geo}}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}_c)$. When $\hat{\boldsymbol{\mu}}$ lies on or near the geodesic from $\boldsymbol{\mu}_0$ to $\boldsymbol{\mu}_1$, we have $\alpha \approx \hat{\theta}_0 + \hat{\theta}_1$, which gives $\cos\alpha \approx \cos(\hat{\theta}_0 + \hat{\theta}_1) = \cos\hat{\theta}_0 \cos\hat{\theta}_1 - \sin\hat{\theta}_0 \sin\hat{\theta}_1$, and hence $\cos A \approx -1$, i.e., $A \approx \pi$.

In our setting, MLLM features have angular standard deviation $\approx 1.7^\circ$ and the dataset size is $N \gtrsim 10^3$, so $d_{\text{geo}}(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) \lesssim 0.1^\circ$. This makes the deviation of A from π negligible, and the inequality $d_{\text{geo}}(\phi(\boldsymbol{\mu}_0), \phi(\boldsymbol{\mu}_1)) \geq d_{\text{geo}}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$ holds in all our experiments with a wide margin.

Remark 3 (Effect on individual data points)**.** *Proposition 1 is stated for the class means, but its practical impact extends to individual data points. For a sample $\mathbf{x} \sim \text{vMF}(\boldsymbol{\mu}_c, \kappa_c)$ with high concentration ($\kappa_c \gg 1$), $\mathbf{x}$ is close to $\boldsymbol{\mu}_c$ and hence $\text{Log}_{\boldsymbol{\mu}}(\mathbf{x})$ points approximately in the direction of $\text{Log}_{\boldsymbol{\mu}}(\boldsymbol{\mu}_c)$, with a small perturbation due to within-class variation. After ℓ_2-normalisation, $\phi(\mathbf{x})$ is concentrated around $\phi(\boldsymbol{\mu}_c)$. Since $\phi(\boldsymbol{\mu}_0)$ and $\phi(\boldsymbol{\mu}_1)$ are antipodal, the two class clusters in the centred space are approximately antipodally opposed, yielding near-maximal between-class separation. The within-class angular spread may change, but the between-class geodesic distance is always expanded. This is confirmed empirically in Figure 3 (main text): after centering, the mean geodesic distance difference $\Delta\mu$ increases from 2.7° to 27.8°.*

A.3 Proof of Proposition 2: Unified Fréchet Mean as Rotational Alignment

We restate the proposition:

Proposition (Proposition 2, restated)**.** *Let the synthetic and real domain distributions be $\text{vMF}(\boldsymbol{\mu}_S, \kappa_S)$ and $\text{vMF}(\boldsymbol{\mu}_R, \kappa_R)$, respectively, where $\boldsymbol{\mu}_R = R\boldsymbol{\mu}_S$ for some rotation $R \in \text{SO}(D)$. Then the unified Fréchet mean $\boldsymbol{\mu}_{\text{unified}}$ of the mixture lies on the geodesic from $\boldsymbol{\mu}_S$ to $\boldsymbol{\mu}_R$. In the symmetric case ($\kappa_S = \kappa_R$ and equal sample sizes), $\boldsymbol{\mu}_{\text{unified}}$ is the geodesic midpoint:*

$$d_{\text{geo}}(\boldsymbol{\mu}_{\text{unified}}, \boldsymbol{\mu}_S) = d_{\text{geo}}(\boldsymbol{\mu}_{\text{unified}}, \boldsymbol{\mu}_R) = \tfrac{1}{2}\, d_{\text{geo}}(\boldsymbol{\mu}_S, \boldsymbol{\mu}_R). \tag{19}$$

This midpoint provides a symmetric reference that simultaneously absorbs the rotational bias from both domains.

Proof. The proof combines the geodesic placement result from Lemma 1 with a reflection symmetry argument.

Step 1: Placement on the geodesic. The two domain distributions are vMF with means $\boldsymbol{\mu}_S$ and $\boldsymbol{\mu}_R$. By Lemma 1 (applied with class $0 = S$, class $1 = R$), the population Fréchet mean of the mixture lies on the geodesic from $\boldsymbol{\mu}_S$ to $\boldsymbol{\mu}_R$:

$$\boldsymbol{\mu}_{\text{unified}} = \text{Slerp}(\boldsymbol{\mu}_S, \boldsymbol{\mu}_R, t_0) \quad \text{for some } t_0 \in (0, 1). \tag{20}$$

Step 2: Midpoint property in the symmetric case. Assume $\kappa_S = \kappa_R \equiv \kappa$ and equal mixture weights $\pi_S = \pi_R = \frac{1}{2}$. Define $\delta = d_{\text{geo}}(\boldsymbol{\mu}_S, \boldsymbol{\mu}_R)$ (the rotational offset, empirically $\approx 5^\circ$ in our setting).

Consider the isometric reflection $\mathcal{R}$ of $\mathcal{S}^{D-1}$ that swaps $\boldsymbol{\mu}_S \leftrightarrow \boldsymbol{\mu}_R$. Geometrically, $\mathcal{R}$ is the geodesic reflection through the midpoint $\boldsymbol{\mu}_{\text{mid}} = \text{Slerp}(\boldsymbol{\mu}_S, \boldsymbol{\mu}_R, \frac{1}{2})$, restricted to the great circle containing

$\boldsymbol{\mu}_S$ and $\boldsymbol{\mu}_R$, and extended to all of $\mathcal{S}^{D-1}$ via the orthogonal reflection in the hyperplane equidistant from $\boldsymbol{\mu}_S$ and $\boldsymbol{\mu}_R$. Formally, let $\mathbf{e}_1 = (\boldsymbol{\mu}_S + \boldsymbol{\mu}_R)/\|\boldsymbol{\mu}_S + \boldsymbol{\mu}_R\|$ and $\mathbf{e}_2 = (\boldsymbol{\mu}_R - \boldsymbol{\mu}_S)/\|\boldsymbol{\mu}_R - \boldsymbol{\mu}_S\|$. Then $\mathcal{R}$ acts as

$$\mathcal{R}(\mathbf{x}) = \mathbf{x} - 2(\mathbf{e}_2^\top \mathbf{x})\, \mathbf{e}_2, \tag{21}$$

which maps $\boldsymbol{\mu}_S \mapsto \boldsymbol{\mu}_R$ and vice versa.

Under $\mathcal{R}$, the distribution $\mathrm{vMF}(\boldsymbol{\mu}_S, \kappa)$ maps to $\mathrm{vMF}(\boldsymbol{\mu}_R, \kappa)$ and vice versa. Since $\kappa_S = \kappa_R$ and $\pi_S = \pi_R$, the mixture distribution is $\mathcal{R}$-invariant. Consequently, $F(\mathcal{R}(\boldsymbol{\mu})) = F(\boldsymbol{\mu})$ for all $\boldsymbol{\mu}$. Since $\boldsymbol{\mu}_{\text{unified}}$ is the unique minimiser, $\mathcal{R}(\boldsymbol{\mu}_{\text{unified}}) = \boldsymbol{\mu}_{\text{unified}}$. The only point on the geodesic from $\boldsymbol{\mu}_S$ to $\boldsymbol{\mu}_R$ that is fixed by $\mathcal{R}$ is the midpoint $\boldsymbol{\mu}_{\text{mid}}$. Therefore $\boldsymbol{\mu}_{\text{unified}} = \boldsymbol{\mu}_{\text{mid}}$, establishing Eq. (19).

Step 3: Symmetric absorption of rotational bias. After centering with $\boldsymbol{\mu}_{\text{unified}} = \boldsymbol{\mu}_{\text{mid}}$ as the base point, the two domains are projected symmetrically into the tangent space:

$$\mathrm{Log}_{\boldsymbol{\mu}_{\text{mid}}}(\boldsymbol{\mu}_S) = -\tfrac{\delta}{2}\,\mathbf{u}, \qquad \mathrm{Log}_{\boldsymbol{\mu}_{\text{mid}}}(\boldsymbol{\mu}_R) = +\tfrac{\delta}{2}\,\mathbf{u}, \tag{22}$$

where $\mathbf{u}$ is the unit tangent vector pointing from $\boldsymbol{\mu}_S$ toward $\boldsymbol{\mu}_R$. Each domain's features are displaced by only $\delta/2$ (half the total offset) from the base point. This is optimal in the minimax sense: it minimises the maximum domain-specific displacement:

$$\boldsymbol{\mu}_{\text{mid}} = \operatorname*{arg\,min}_{\boldsymbol{\mu} \in \mathcal{S}^{D-1}} \max\!\big\{ d_{\text{geo}}(\boldsymbol{\mu}, \boldsymbol{\mu}_S),\ d_{\text{geo}}(\boldsymbol{\mu}, \boldsymbol{\mu}_R) \big\}.$$

By contrast, using either domain-specific mean as the base point would leave a residual bias of $\delta \approx 5^\circ$ for the other domain, causing systematic misalignment. □

Remark 4 (Asymmetric case)**.** *When $\kappa_S \neq \kappa_R$ or $\pi_S \neq \pi_R$, the unified Fréchet mean still lies on the geodesic from $\boldsymbol{\mu}_S$ to $\boldsymbol{\mu}_R$ (by Lemma 1), but is biased toward the domain with larger $\pi_c\kappa_c$ (more samples or higher concentration). In our setting, $|\mathcal{D}_{\text{syn}}| \approx 2{,}000$ and $|\mathcal{D}_{\text{real}}| \approx 3{,}912$–$97{,}396$, so $\boldsymbol{\mu}_{\text{unified}}$ is naturally closer to $\boldsymbol{\mu}_R$. This is desirable: the real (test) domain receives a smaller centering correction, preserving its features' fine structure, while the synthetic domain absorbs a larger share of the alignment adjustment.*

Remark 5 (Centering absorbs domain bias for individual features)**.** *For an individual synthetic feature $\tilde{f}_{\text{syn}}$ close to $\boldsymbol{\mu}_S$ and a real feature $\tilde{f}_{\text{real}}$ close to $\boldsymbol{\mu}_R$, the raw cosine similarity is corrupted by the $\delta \approx 5^\circ$ offset: $\tilde{f}_{\text{syn}}^\top \tilde{f}_{\text{real}}$ underestimates the semantic similarity. After centering, both features are expressed in the tangent-space coordinate system at $\boldsymbol{\mu}_{\text{unified}}$, and the systematic offset is absorbed. The centred features $\hat{f}_{\text{syn}}$ and $\hat{f}_{\text{real}}$ are directly comparable, as confirmed by the $+5.78\%$ AUC improvement on UCF-Crime (Table 2, M0a→M0b).*

B Synthetic Calibration Dataset Details

Zero-shot declaration. The synthetic calibration set $\mathcal{D}_{\text{syn}}$ is constructed *entirely independently* of every evaluation benchmark. No image, frame, video, scene description, or anomaly-category taxonomy from XD-Violence, UCF-Crime, or UBnormal is used at any stage of the generation pipeline. The only supervision signal is a *binary* label $y \in \{\texttt{normal}, \texttt{abnormal}\}$; fine-grained anomaly categories (e.g., "Personal Emergency," "Public Misconduct") are used solely to promote visual diversity during generation and are **never** exposed to the SphereVAD inference pipeline. Consequently, SphereVAD satisfies the strict zero-shot criterion: the model has access to neither target-domain data nor target-domain category definitions.

B.1 Generation Pipeline

The synthetic data are produced by a fully automated **two-stage** pipeline (Figure 5):

Stage 1: Structured scene description generation (LLM). A large language model (Claude Sonnet 4.5, accessed via API) receives a carefully designed meta-prompt specifying six broad anomaly meta-categories—*Violent Conflict*, *Crime*, *Traffic Accident*, *Personal Emergency*, *Environmental Incident*, and *Public Misconduct*—together with detailed structural constraints. For each meta-category the LLM generates multiple *scene pairs*, where every pair consists of:

- **Shared context** P_c: a specific location, time of day, and environmental details (e.g., "Dusk, underground parking garage, wet concrete floor").
- **Shared characters** P_{char}: number of people, appearance, and clothing, kept *identical* across all frames and both branches.
- **Abnormal branch**: a four-frame storyboard $(F_1^{\text{pre}}, F_2^{\text{start}}, F_3^{\text{peak}}, F_4^{\text{post}})$ depicting the temporal evolution of an anomalous event.
- **Normal branch**: a matching four-frame storyboard under the same scene and characters but depicting routine, non-anomalous behaviour.

The meta-prompt enforces several quality constraints: (i) F_1^{pre} must be nearly identical across the two branches to ensure a shared baseline; (ii) F_3^{peak} must contain at least three distinct, visually verifiable anomalous cues in the abnormal branch; (iii) character clothing and scene background must remain strictly consistent across all four frames; (iv) no depictions of minors are permitted. The LLM outputs each pair as a structured JSON object. Each generation call produces $G{=}15$ pairs; $R{=}2$ rounds are executed per meta-category, and additional rounds are run with varied random seeds until the total reaches approximately 1,000 pairs.

Stage 2: Image synthesis (text-to-image model). Each branch's four-frame description is converted into a single prompt by prepending the shared-context and character descriptions and appending a style prefix:

```
realistic photography, surveillance camera footage, photorealistic,
raw photo, 2x2 grid layout, four sequential frames, left to right
top to bottom.
```

The composed prompt is sent to Gemini 2.5 Flash (image generation mode, accessed via API), which returns a single square image containing a $2{\times}2$ grid of the four frames (arranged left-to-right, top-to-bottom: F_1, F_2, F_3, F_4). Retry logic (up to 2 retries per image) and timeout handling ensure robustness; permanently failed pairs (exceeding 3 non-timeout failures) are discarded.

Design motivation for the $2{\times}2$ grid. Encoding four temporal frames in a single image serves three purposes. First, it substantially reduces generation cost and latency: producing one composite image per branch rather than four separate images cuts API calls by $4\times$, making the pipeline practical at scale. Second, it improves inter-frame visual consistency: because the text-to-image model generates all four sub-frames jointly within a single image, character appearance, scene layout, and lighting remain naturally coherent across the temporal sequence—an outcome that is difficult to guarantee when frames are generated independently. Third, it ensures structural alignment with the test-time input format: during inference, each test clip is likewise represented as a $2{\times}2$ grid of four consecutive sampled frames, processed with the same prompt template and fed to the MLLM as a single composite image. By matching the spatial layout, aspect ratio, and prompt structure between calibration and test inputs, the intermediate-layer features extracted from both domains occupy the same representational subspace, which is critical for the unified Fréchet mean centering (Section 3.2) to effectively absorb domain bias. At inference time, the MLLM processes each composite image as a single input; the $2{\times}2$ spatial layout implicitly conveys temporal progression ($F_1 \to F_2 \to F_3 \to F_4$), enabling the model to capture temporal dynamics through its intermediate-layer features without requiring explicit video-level input.

B.2 Data Format & Pairing Strategy

Pairing index. All generated pairs are tracked in a unified CSV file (`pair_index.csv`) with the schema shown in Table 5. Each row links one normal image and one abnormal image that share exactly the same scene and characters.

"Same background, different event" design. The paired design is central to directional prototype calibration. Because each normal–abnormal pair shares the same scene context and characters, the discriminative direction between the two corresponding features in MLLM intermediate-layer space reflects *semantic* differences (routine vs. anomalous event) rather than *nuisance* differences (lighting, scene layout, character appearance). Spherical K-Means clustering on these paired features therefore yields prototypes whose mean directions capture anomaly semantics with minimal scene-specific confounds.

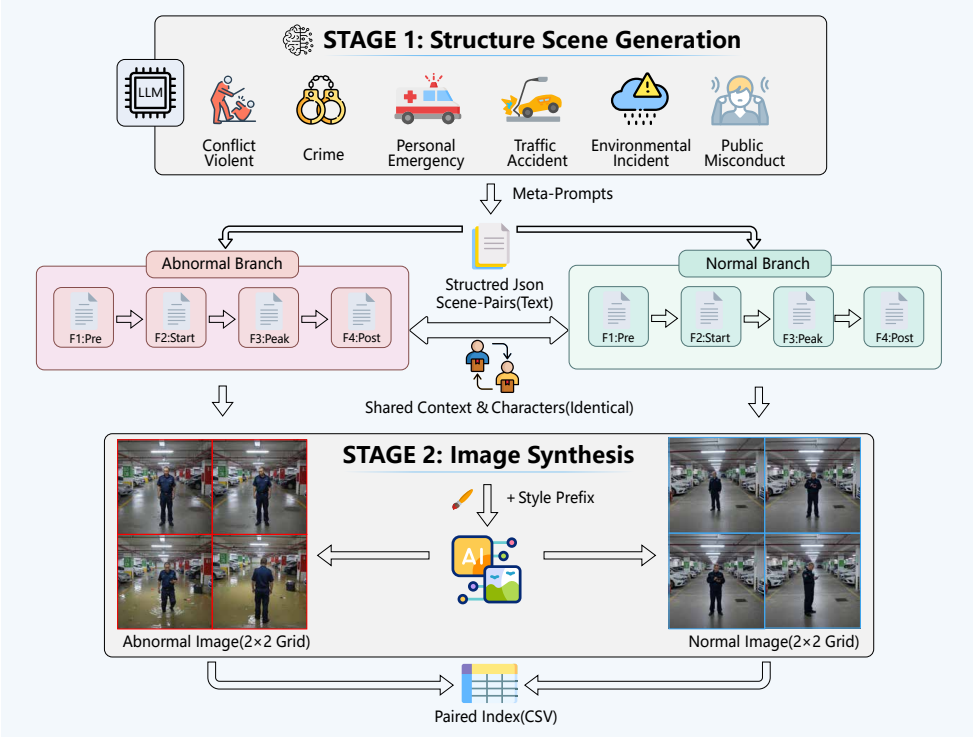


Figure 5: Two-stage synthetic calibration data generation pipeline. **Stage 1:** Claude Sonnet 4.5 generates structured JSON scene-pair descriptions with shared context, shared characters, and four-frame storyboards for both abnormal and normal branches. **Stage 2:** Gemini 2.5 Flash renders each branch as a 2×2 grid image containing four temporally ordered frames.

Table 5: Schema and example entries of `pair_index.csv`.

pair_id	source_label	normal_path	abnormal_path
Violent Conflict_01	Violent Conflict	.../normal/VC_0001.jpg	.../abnormal/VC_0002.jpg
Crime_05	Crime	.../normal/Cr_0009.jpg	.../abnormal/Cr_0010.jpg
Traffic Accident_12	Traffic Accident	.../normal/TA_0023.jpg	.../abnormal/TA_0024.jpg
Personal Emergency_03	Personal Emergency	.../normal/PE_0005.jpg	.../abnormal/PE_0006.jpg
Public Misconduct_08	Public Misconduct	.../normal/PM_0015.jpg	.../abnormal/PM_0016.jpg

Directory structure. The dataset is organised hierarchically:

```
Syn-4img/
|-- pair_index.csv
|-- Violent Conflict/
|   |-- abnormal/   VC_0002.jpg, VC_0004.jpg, ...
|   |-- normal/     VC_0001.jpg, VC_0003.jpg, ...
|   |-- Violent Conflict.csv          (success log)
|   |-- Violent Conflict_failed.csv   (failure log)
|-- Crime/
|   |-- abnormal/   ...
|   |-- normal/     ...
...
```

Per-label success and failure CSVs enable deterministic checkpoint–resume: the pipeline can be interrupted and restarted without regenerating existing images.

B.3 Sample Images

Figure 6 shows representative normal–abnormal paired examples from four of the six meta-categories. Each row displays one pair: the left column shows the generated 2×2 grid image (containing four temporally ordered frames F_1–F_4, arranged left-to-right, top-to-bottom), the centre column shows the corresponding LLM-generated four-frame storyboard description, and the right column shows the counterpart branch image from the same pair. Within each pair, the top row is the **normal** branch and the bottom row is the **abnormal** branch. Note that within each pair the scene layout, character appearance, and camera viewpoint are visually consistent; only the depicted behaviour and the storyboard descriptions differ.

B.4 Sample Size Justification

The final calibration set comprises approximately **2,000 images** ($\sim$1,000 normal–abnormal pairs), distributed roughly uniformly across the six meta-categories ($\sim$160–180 pairs each). We justify the sufficiency of this modest sample size from three perspectives.

(1) Precedent: minimal samples activate latent MLLM discriminability. HiProbe [7] demonstrates that as few as 10–50 labelled samples suffice to probe and activate the discriminative capacity already latent in MLLM intermediate layers; the primary role of the calibration data is to *locate* the relevant directions in feature space, not to *learn* new representations. SphereVAD follows the same principle: synthetic images serve as *directional calibrators* rather than training data.

(2) vMF parameter estimation requires only the mean direction. The vMF distribution $\mathrm{vMF}(\mathbf{x};\boldsymbol{\mu},\kappa) \propto \exp(\kappa\,\boldsymbol{\mu}^\top\mathbf{x})$ is parameterised by its mean direction $\boldsymbol{\mu}$ and concentration κ. Spherical K-Means estimates $\boldsymbol{\mu}$ as the ℓ_2-normalised cluster centroid—a sample mean that converges at rate $O(N^{-1/2})$ on the sphere [33]. With $\sim$50–100 samples per cluster across the partition configurations, the angular estimation error of $\boldsymbol{\mu}$ is well below 1°, which is negligible compared with the inter-class separation of 27.8° achieved after spherical centering (Figure 3). The concentration parameter κ is calibrated as a global hyperparameter rather than estimated per-cluster, further reducing sample requirements.

(3) No content overlap with any test benchmark. All images are procedurally generated from LLM-authored scene descriptions followed by text-to-image synthesis. No real surveillance footage, no frames from XD-Violence, UCF-Crime, or UBnormal, and no benchmark-specific scene or category metadata are used at any point. The six meta-categories (Violent Conflict, Crime, Traffic Accident, Personal Emergency, Environmental Incident, Public Misconduct) are broad semantic anchors designed to cover the generic space of "anomalous events in public settings"; they do not correspond one-to-one to the category taxonomies of any benchmark (e.g., UCF-Crime defines 13 categories, XD-Violence defines 6 partially overlapping categories). This complete independence guarantees that the zero-shot evaluation protocol is not compromised.

B.5 Sub-category Taxonomy and Label Independence

To promote visual diversity within each meta-category, the LLM is instructed to generate scene descriptions spanning a variety of *sub-categories*. These sub-categories serve exclusively as diversity-inducing prompts for the text-to-image model and are **never** exposed to the SphereVAD inference pipeline—only the binary label $y \in \{\texttt{normal}, \texttt{abnormal}\}$ is retained. Table 6 enumerates all 34 sub-categories used across the six meta-categories.

Label independence from evaluation benchmarks. We explicitly verify that none of the 34 sub-category names coincides with any category label defined in the three evaluation benchmarks. Table 7 presents a systematic comparison.

Three key observations confirm label independence:

1. **No exact string match:** None of the 34 synthetic sub-category names appears verbatim in any benchmark's category list. For instance, UCF-Crime uses "Vandalism" whereas our taxonomy uses "Graffiti"; UCF-Crime uses "RoadAccidents" (a broad category) whereas we use the finer-grained "Hit and Run," "Pedestrian Hit," and "Bicycle Accident."



Figure 6: Example paired synthetic calibration data from four meta-categories. Each meta-category occupies two rows: the top row is the **normal** branch and the bottom row is the **abnormal** branch. For each row, the left image shows the generated 2×2 grid (four sequential frames F_1–F_4), the centre text shows the LLM-generated storyboard description for each frame, and the right image shows the corresponding counterpart grid from the same scene pair. From top to bottom: *Environmental Incident*, *Personal Emergency*, *Public Misconduct*, *Crime*. Within each pair, the scene, characters, and camera viewpoint are shared; only the depicted event differs.

Table 6: Complete list of the 34 sub-categories used to diversify synthetic scene generation, grouped by meta-category. These labels are used only during prompt composition for data generation and are discarded before inference.

Meta-category	**Sub-categories**
Violent Conflict	Stabbing, Hostage Situation, Chasing
Crime	Pickpocketing, Vehicle Theft, Trespassing, Property Damage
Traffic Accident	Bicycle Accident, Hit and Run, Pedestrian Hit, Running Red Light, Road Obstruction
Personal Emergency	Drowning, Electrocution, Fainting, Fall Down, Trapped
Environmental Incident	Abandoned Object, Building Collapse, Falling Object, Flooding, Gas Leak, Smoke Emission
Public Misconduct	Animal Attack, Climbing Structure, Crowd Panic, Crowd Stampede, Drunk Behavior, Fence Jumping, Graffiti, Illegal Dumping, Jaywalking (Dangerous), Loitering (Suspicious), Substance Abuse

Table 7: Category taxonomies of the three evaluation benchmarks versus the synthetic sub-categories. No exact label match exists between any benchmark category and any synthetic sub-category. Semantically related pairs (e.g., UCF-Crime "RoadAccidents" vs. our "Hit and Run") differ in granularity and definition; crucially, sub-category labels are never used during inference.

Benchmark	**Benchmark categories**	**Overlap with synthetic sub-cats**
UCF-Crime (13 categories)	Abuse, Arrest, Arson, Assault, Burglary, Explosion, Fighting, RoadAccidents, Robbery, Shooting, Shoplifting, Stealing, Vandalism	**No exact match.** Semantic proximity exists for a few pairs (e.g., "RoadAccidents" $\leftrightarrow$ "Hit and Run" / "Pedestrian Hit"; "Stealing" $\leftrightarrow$ "Vehicle Theft" / "Pickpocketing"; "Vandalism" $\leftrightarrow$ "Graffiti"), but names, granularity, and visual scope differ.
XD-Violence (6 categories)	Abuse, Car Accident, Explosion, Fighting, Riot, Shooting	**No exact match.** "Car Accident" is semantically related to "Hit and Run" / "Bicycle Accident," but the synthetic labels are finer-grained and do not include car-to-car collisions.
UBnormal (pose/motion anomalies)	Defined by abnormal pose / motion patterns (e.g., falling, lying down, running); no named crime categories.	**No overlap.** UBnormal does not define named anomaly categories comparable to our sub-categories.

2. **Deliberate granularity mismatch:** The synthetic sub-categories are intentionally defined at a different level of granularity than benchmark categories. This design choice ensures that even semantically proximate concepts (e.g., "Stealing" vs. "Pickpocketing") describe distinct visual scenarios, preventing implicit information leakage about benchmark-specific category boundaries.
3. **Labels discarded before inference:** Most importantly, all sub-category labels are used *exclusively* during the data generation stage (Stage 1, Section B.1) to instruct the LLM to produce diverse scene descriptions. They are completely stripped from the data before any feature extraction or prototype estimation occurs. The SphereVAD inference pipeline receives only the binary normal/abnormal label and the composite grid image, making it categorically impossible for sub-category information to influence anomaly scores.

C Feature Extraction

This section details the complete feature extraction pipeline of SphereVAD. Section C.1 presents the full prompt template and its design rationale. Section C.2 explains why the prompt structure must respect causal attention constraints. Section C.3 describes the Dynamic Layer Saliency Probing (DLSP) procedure used to select the optimal extraction layer, including comparative curves across

four MLLM backbones. Section C.4 details the dual-stream feature design and the complementary roles of the main feature f^l and the visual feature f^v.

C.1 Complete Prompt Template

The prompt fed to the frozen MLLM follows a **Text – Image – Text** sandwich structure. Each input sample (whether a synthetic calibration image or a test video clip) is represented as four temporally ordered sub-images $\{I_1, I_2, I_3, I_4\}$. For synthetic data these are obtained by splitting a 2×2 grid image (left-to-right, top-to-bottom); for test videos they are four uniformly sampled frames from the clip. When `RESIZE_SUBIMAGES` is enabled, each sub-image is resized to 336×336 pixels via Lanczos interpolation to ensure uniform spatial resolution across all inputs. The full prompt is assembled as follows:

> `[Part 1 -- Task Instruction]`
>
> You are a professional video security analysis assistant. The following four consecutive video frames record the temporal evolution of the same scene.
>
> [Anomaly Whitelist]
> Anomalous events are limited to the following 6 categories:
> [Violent Conflict], [Crime], [Traffic Accident], [Personal Emergency], [Environmental Incident], [Public Misconduct].
>
> With the above classification criteria in mind, carefully observe the following frames to determine whether a matching anomalous event is present:
>
> `[Sub-image` I_1`]` `[Sub-image` I_2`]` `[Sub-image` I_3`]` `[Sub-image` I_4`]`
>
> `[Part 2 -- Output Format]`
>
> Based on the above frames, strictly follow the 4-step output format below (always start with 'Yes' or 'No'):
>
> 1. Final determination: [Yes or No].
> 2. Anomaly category match: [Format: Category – Specific sub-label. If No, output: None].
> 3. Spatiotemporal action description: [Briefly describe character interactions, action continuity, and object state changes over time].
> 4. Confidence assessment: [High / Medium / Low. If category is 'None', output: None].

The prompt is wrapped as a single-turn chat message with `role=user` and processed through the model's chat template via `apply_chat_template` (with `add_generation_prompt=True` and `enable_thinking=False`). The processor then jointly tokenises the text and encodes the images, yielding the final input tensors (`input_ids`, `attention_mask`, `pixel_values`, `image_grid_thw`, etc.).

Design rationale. The prompt is designed with four considerations:

1. **Structured anomaly whitelist.** The six meta-categories (Violent Conflict, Crime, Traffic Accident, Personal Emergency, Environmental Incident, Public Misconduct) are explicitly enumerated in Part 1 to prime the model's internal representations toward anomaly-relevant semantics. These categories are intentionally broad and do not correspond one-to-one to the category taxonomies of any benchmark (e.g., UCF-Crime defines 13 categories, XD-Violence defines 6 partially overlapping ones). The whitelist serves as a *semantic anchor* that activates latent anomaly-discriminative directions in the intermediate layers, without leaking benchmark-specific category information.
2. **Temporal framing.** The phrase "four consecutive video frames record the temporal evolution of the same scene" cues the model to interpret the four sub-images as a temporal sequence rather than four independent observations. This encourages the intermediate-layer features to encode temporal dynamics (e.g., action progression, state changes) rather than merely static scene appearance.
3. **Structured output format in Part 2.** Although SphereVAD *never* decodes the model's textual output, the structured output instructions in Part 2 are critical. Through the causal attention mechanism (Section C.2), the generation prompt tokens attend to both the visual tokens and the task instructions, producing hidden states that integrate anomaly-category reasoning, spatiotemporal

action analysis, and confidence estimation. The last-token hidden state at the extraction layer thus encodes a rich fusion of these reasoning pathways.

4. **Consistent format across synthetic and real inputs.** Both synthetic calibration images (split from 2×2 grids) and real test clips (four sampled frames) are processed with *identical* prompt templates and image preprocessing. This format consistency ensures that the intermediate-layer features from both domains occupy the same representational subspace, which is essential for the unified Fréchet mean centering (Section 3.2) to effectively absorb domain bias.

C.2 Causal Attention Justification

Modern MLLMs (including the Qwen-VL family, InternVL, and LLaVA) employ **causal (autoregressive) attention**: each token can attend only to itself and all preceding tokens in the sequence. This architectural constraint has a direct and critical impact on which token positions carry the richest fused representations.

Why the visual-last token alone is insufficient. Under causal attention, vision tokens can attend to all preceding tokens—including the Part 1 task instruction—but *cannot* attend to any tokens that follow them, including the Part 2 output format instructions and the generation prompt. Consequently, the hidden state of the last vision token (f^v, "visual-last") encodes the visual content contextualised by the task instruction, but lacks the reasoning structure imposed by the output format. If one were to extract *only* f^v for anomaly scoring, the feature would capture scene appearance but miss the anomaly-category reasoning and confidence estimation cues induced by Part 2.

Why Part 1 text must precede vision tokens. Placing the task instruction *before* the vision tokens ensures that every vision token can attend to the anomaly whitelist and temporal framing cues during the forward pass. This contextualises the visual representations from the earliest processing stage: the model does not merely "see" the images but "sees them through the lens of anomaly detection." If the task instruction were placed *after* the vision tokens, the visual representations would be computed without any task-specific context, and only the later text tokens would benefit from cross-modal attention—severely limiting the discriminative quality of the visual features.

Why the last sequence token carries the richest information. The last token in the full sequence (i.e., the final token of the generation prompt, after Part 2) is the only position that can attend to *all* preceding tokens: Part 1 instruction $\rightarrow$ vision tokens $\rightarrow$ Part 2 output format $\rightarrow$ generation prompt. Its hidden state at the extraction layer therefore integrates:

- **Task semantics** from Part 1 (anomaly categories, temporal framing);
- **Visual content** from the four sub-image tokens;
- **Reasoning structure** from Part 2 (determination, category matching, spatiotemporal description, confidence).

This is why we designate the last-token hidden state as the *main feature* f^l, used for anomaly scoring, while the visual-last hidden state f^v serves a complementary role as a *scene descriptor* for cross-video attention (HSA) and intra-video neighbour consensus (SGP), where scene appearance similarity rather than anomaly reasoning is the relevant signal.

Token sequence structure. The MLLM processes tokens in the order: system prompt $\rightarrow$ Part 1 task instruction $\rightarrow$ vision tokens for sub-images I_1–I_4 $\rightarrow$ Part 2 output format $\rightarrow$ generation prompt (last token). Under causal masking, the last token has the maximal receptive field attending to the entire sequence, while the visual-last token (last vision token of sub-image 4) attends only to Part 1 and all vision tokens but not Part 2.

Vision token localisation. To extract f^v, we must identify the position of the last vision token in the tokenised sequence. The extraction code searches for vision boundary tokens in the following priority order:

1. **Boundary-delimited:** Locate `<|vision_start|>` / `<|vision_end|>` (or `<|img_start|>` / `<|img_end|>`) token pairs. For $K=4$ sub-images, the code identifies the first K such pairs and marks all tokens between each pair as vision tokens. The position of the last marked token is the visual-last position.

2. **Pad-token fallback:** If boundary tokens are not found (e.g., in models using a different tokenisation scheme), the code searches for <|image_pad|>, <|vision_pad|>, or <|img_pad|> tokens and takes the last occurrence.

If neither method succeeds (an edge case that did not occur in our experiments), f^v falls back to f^l (the last-token feature).

C.3 Layer Selection via DLSP

Motivation. Not all intermediate layers of a pre-trained MLLM carry equally discriminative anomaly information. Early layers tend to encode low-level visual patterns (edges, textures), while the final layers are optimised for language generation and may discard fine-grained visual semantics. The optimal extraction layer—where anomaly-discriminative information peaks—must be identified empirically. We adopt **Dynamic Layer Saliency Probing (DLSP)**, inspired by HiProbe [7], as a lightweight, *training-free* method to locate this layer.

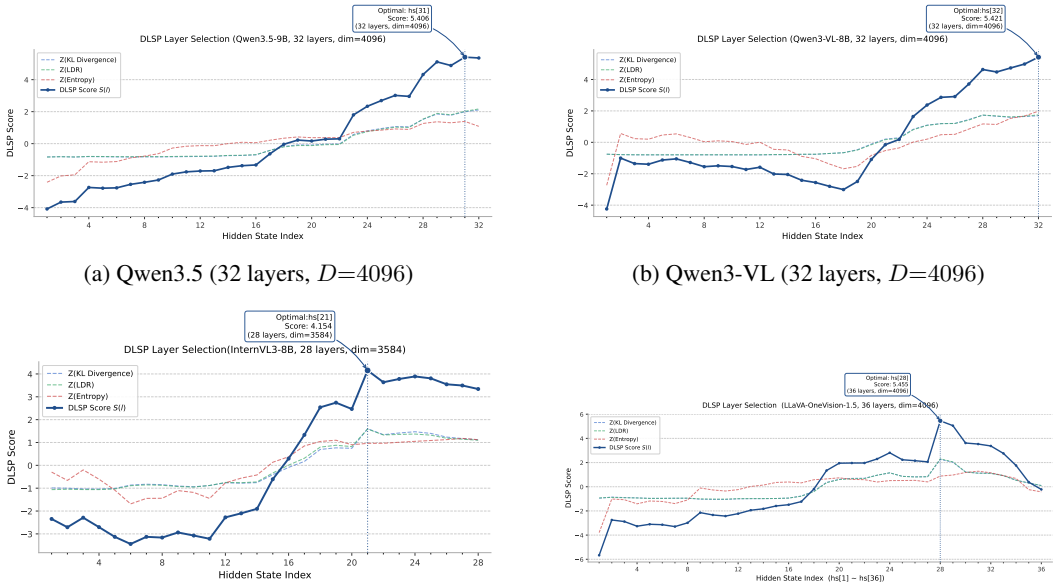
DLSP procedure. DLSP evaluates each candidate layer $\ell \in \{1, 2, \ldots, L\}$ (where L is the total number of hidden-state layers) by measuring how well the last-token features at that layer separate the normal and anomalous classes in the synthetic calibration set $\mathcal{D}_{\text{syn}}$. The procedure computes three complementary separability metrics at each layer and fuses them into a single composite score:

1. **Extract layer-wise features.** For each synthetic sample $x_j \in \mathcal{D}_{\text{syn}}$, extract the last-token hidden state at every layer ℓ, yielding $\{f_j^{(\ell)}\}_{\ell=1}^{L}$. All features are ℓ_2-normalised onto $\mathcal{S}^{D-1}$.
2. **Compute per-layer separability metrics.** For each layer ℓ, three metrics are computed between the normal and anomalous feature distributions:
   - **KL Divergence:** measures the distributional divergence between the cosine similarity distributions of normal and anomalous features to their respective class centroids.
   - **Log Density Ratio (LDR):** quantifies the separation of per-class density estimates in the feature space.
   - **Entropy:** captures the predictability of class membership from the feature representations (lower entropy indicates better separability).

   Each metric is standardised to a Z-score across layers to ensure commensurability.
3. **Composite saliency score.** The three Z-scored metrics are combined into a single DLSP saliency score:
$$\text{DLSP}(\ell) = Z_{\text{KL}}^{(\ell)} + Z_{\text{LDR}}^{(\ell)} - Z_{\text{Entropy}}^{(\ell)}, \tag{23}$$
   where the entropy term is subtracted because lower entropy corresponds to better separability. A higher composite score indicates that features at layer ℓ exhibit stronger anomaly discriminability across all three criteria.
4. **Select optimal layer.** The extraction layer is chosen as $\ell^* = \arg\max_\ell \text{DLSP}(\ell)$.

Key properties. DLSP is computed entirely on the synthetic calibration set and requires no access to test data or their labels. The procedure involves only forward passes (no gradient computation) and lightweight statistical calculations, making it computationally negligible (<1 minute for all layers on a single GPU). Once ℓ^* is determined, it is fixed for all subsequent inference.

Results across four backbones. Figure 7 plots the DLSP saliency curves for all four MLLM backbones evaluated in our study. The four architectures vary in depth and representation dimensionality: Qwen3.5 has 32 layers with D=4096, Qwen3-VL has 32 layers with D=4096, LLaVA-OneVision-1.5 has 36 layers with D=4096, and InternVL3 has 28 layers with D=3584.



(a) Qwen3.5 (32 layers, D=4096)

(b) Qwen3-VL (32 layers, D=4096)

(c) InternVL3 (28 layers, D=3584)

(d) LLaVA-OneVision-1.5 (36 layers, D=4096)

Figure 7: **DLSP saliency curves for four MLLM backbones.** Each panel displays the per-layer composite score $S(\ell)$ (solid curve, Eq. (23)) alongside the three normalized components: $Z(\text{KL Divergence})$, $Z(\text{LDR})$, and $Z(\text{Entropy})$ (dashed curves). Vertical dashed lines indicate the selected feature extraction layer ℓ^*. **(a)** Qwen3.5 peaks at layer index 31 (score 5.406). **(b)** Qwen3-VL peaks at layer index 31 (score 5.421). **(c)** InternVL3 peaks at layer index 21 (score 4.154). **(d)** LLaVA-OneVision-1.5 peaks at layer index 28 (score 5.455). Across all architectures, composite saliency increases through early representation stages and isolates an optimal deep layer prior to language decoding heads.

Several consistent patterns emerge:

- **Deep intermediate saliency peaks.** Across all evaluated architectures, anomaly discriminability peaks within intermediate-to-deep layers rather than initial visual projections or output token prediction heads. This confirms that semantic divergence between normal and abnormal events is maximally resolved in intermediate latent representations.
- **Optimal extraction layer profiles.** The layer yielding maximum DLSP saliency varies with backbone architecture: Qwen3.5 peaks at layer index 31 (D=4096, DLSP^*=5.406), Qwen3-VL at layer index 31 (D=4096, DLSP^*=5.421), LLaVA-OneVision-1.5 at layer index 28 (D=4096, DLSP^*=5.455), and InternVL3 at layer index 21 (D=3584, DLSP^*=4.154). Across these architectures, the identified layers generally occupy the upper quartile range ($[0.75L,\ 0.97L]$) of network depth.
- **Complementarity of geometric metrics.** As illustrated by the constituent curves (dashed lines in Figure 7), individual indicators (KL Divergence, LDR, and Entropy) achieve their individual maxima at disparate network depths. The composite DLSP formulation balances these complementary geometric properties, providing a robust heuristic for intra-backbone layer determination without downstream supervision.
- **Intra-backbone selection vs. cross-backbone transfer.** While DLSP reliably identifies the optimal layer index ℓ^* for any given architecture, the absolute magnitude of DLSP^* does not directly correlate with cross-backbone downstream performance (Table 3). For instance, LLaVA-OneVision-1.5 yields the highest composite score (5.455) despite lower downstream metrics than InternVL3 and Qwen variants. This confirms that DLSP operates effectively as an internal layer-selection criterion on the spherical manifold, whereas absolute benchmark accuracy remains governed by backbone-level visual encoders and pre-training distributions.

Sensitivity to layer choice. We further evaluate the robustness of the layer selection by measuring downstream performance (XD-Violence AP) when the extraction layer deviates from ℓ^* by $\pm$1–3

layers. As shown in Table 8, performance degrades gracefully within a ± 2-layer window ($<1\%$ AP drop), confirming that DLSP reliably identifies a high-quality extraction region rather than a fragile single-layer optimum.

Table 8: **Layer sensitivity analysis (Qwen3.5, XD-Violence).** Performance variation when the extraction layer deviates from the DLSP-optimal hidden state 31.

Layer ℓ	28	29	30	**31**	32
AP(%)	85.73	86.42	86.81	**86.99**	86.87
Δ vs. ℓ^*	-1.26	-0.57	-0.18	—	-0.12

C.4 Dual Feature Design

SphereVAD extracts two complementary feature streams from each input, both from the same forward pass of the frozen MLLM:

1. **Main feature $f^l \in \mathbb{R}^D$ (last-token hidden state at layer ℓ^*).** This is the hidden state of the *last sequence token* (i.e., the final token of the generation prompt) at the DLSP-selected layer ℓ^*. As argued in Section C.2, under causal attention the last token has the maximal receptive field, attending to the task instruction, all vision tokens, and the output format. Its hidden state therefore encodes a rich fusion of visual content and anomaly-reasoning structure. f^l is the primary feature used for vMF prototype construction and anomaly scoring.
2. **Visual feature $f^v \in \mathbb{R}^D$ (visual-last hidden state at layer ℓ^*).** This is the hidden state of the *last vision token* (the last token within the vision token span of the fourth sub-image) at the same layer ℓ^*. It captures scene appearance—spatial layout, lighting, object arrangement—contextualised by the Part 1 task instruction but *without* the output-format reasoning of Part 2. f^v is used exclusively for scene-similarity computation in HSA (Section 3.4) and neighbour consensus in SGP (Section 3.5), where the relevant signal is "do these two clips depict visually similar scenes?" rather than "is this clip anomalous?"

Why two features rather than one? Using a single feature for both scoring and scene attention would conflate two distinct objectives. The main feature f^l is optimised (via prompt design and layer selection) to maximise *anomaly discriminability*: normal and anomalous clips in the same scene should have maximally different f^l vectors. The visual feature f^v, by contrast, should maximise *scene similarity*: clips from the same or visually similar scenes should have similar f^v vectors *regardless of whether they are normal or anomalous*. These two objectives are inherently in tension: a feature that is highly anomaly-discriminative will by definition differ between normal and anomalous clips of the same scene, making it a poor scene-similarity measure. Empirically, using f^v for HSA scene attention and f^l for scoring outperforms using f^l for both purposes by $+1.43\%$ AP on XD-Violence, confirming the complementary value of the dual-feature design.

Layer index for visual features. Both f^l and f^v are extracted from the *same* layer ℓ^* (selected by DLSP on the main feature). We experimented with selecting a separate optimal layer for f^v but found no significant improvement ($<0.2\%$ AP on XD), likely because the DLSP-selected layer already represents a good trade-off between visual richness and semantic abstraction. Using the same layer simplifies the pipeline (a single forward pass with a single extraction point) and avoids introducing additional hyperparameters.

Visual feature centering. As described in Section 3.2 of the main text, the main features $\hat{f}^l$ are centered using a unified Fréchet mean computed from *both* synthetic and real features (Eq. (1)), which absorbs the systematic rotational bias between domains. The visual features $\hat{f}^v$ are centered with a *separate* Fréchet mean $\boldsymbol{\mu}^v_{\text{unified}}$ computed from the **test set only**:

$$\boldsymbol{\mu}^v_{\text{unified}} = \text{FréchetMean}\big(\{\tilde{f}^v_{\text{real}}\}\big). \tag{24}$$

The rationale is that visual features serve only for scene-similarity computation among test clips; they are never compared against synthetic prototypes. Therefore, cross-domain alignment is unnecessary,

and including synthetic visual features in the Fréchet mean computation would introduce irrelevant bias from the synthetic scene distribution (which is procedurally generated and does not reflect real surveillance environments).

Feature dimensions and storage. The hidden dimension varies across backbones: D=4096 for Qwen3.5, Qwen3-VL, and LLaVA-OneVision-1.5, and D=3584 for InternVL3. Each clip produces two float32 vectors, totalling $2 \times D \times 4$ bytes per clip (e.g., 32 KB for D=4096). For the largest dataset (XD-Violence test set, ∼97,000 clips), the total feature storage is approximately 3.19 GB (with D=4096). Features are saved as per-video `.pt` files, each containing tensors of shape $[T_i, D]$ for both streams, where T_i is the number of clips in video V_i.

Multi-GPU extraction. Feature extraction is parallelised across multiple GPUs using a producer–consumer architecture. Tasks (synthetic pairs for calibration; video clips for test sets) are placed in a shared queue and distributed to worker processes, each running on a designated GPU. An asynchronous prefetch pool (3 threads per worker) pre-tokenises upcoming inputs while the current input is being processed by the model, effectively hiding I/O and preprocessing latency. For extremely long videos (e.g., >1,000 clips), the clips are split into chunks assigned to different workers and merged after extraction. On 4×A100 GPUs, the complete feature extraction for all three benchmarks plus the synthetic calibration set takes approximately 109 minutes (109.1 min total wall-clock time across subsets).

D SGP: Full Algorithm

This section provides the complete mathematical derivation and algorithmic details of the vMF-guided Spherical Geodesic Pulling (SGP) mechanism described in Section 3.5 of the main text. SGP is applied independently to each video and comprises four stages: MAD-based adaptive tri-classification(§D.1), dominant prototype voting (§D.2), neighbour consensus assignment (§D.3), and score-adaptive SLERP pulling (§D.4). The complete algorithm is given in §D.5.

Throughout this appendix, we consider a single video V with T clips. The enhanced main features $\{\hat{f}_i^{l,\text{enh}}\}_{i=1}^T \subset \mathcal{S}^{D-1}$, visual features $\{\hat{f}_i^v\}_{i=1}^T \subset \mathcal{S}^{D-1}$, and initial vMF anomaly scores $\{s_i^{\text{init}}\}_{i=1}^T \subset [0,1]$ are all given from the preceding HSA stage (Section 3.4).

D.1 MAD-Based Adaptive Ambiguity Interval

A fixed decision threshold (e.g., $s = 0.5$) is inadequate for real-world surveillance because the initial score distribution varies significantly across videos: some videos exhibit a clear bimodal separation between normal and anomalous clips, while others produce scores concentrated near the decision boundary. SGP addresses this via an adaptive ambiguity interval $[\rho_{\text{low}}, \rho_{\text{high}}]$ derived from the Median Absolute Deviation (MAD) [23], a robust dispersion estimator that is resistant to outliers.

Step 1: MAD computation. Let $\tilde{s} = \text{median}(\{s_i^{\text{init}}\}_{i=1}^T)$ denote the median initial score of the video. The MAD is defined as

$$\text{MAD} = \text{median}\big(\{|s_i^{\text{init}} - \tilde{s}|\}_{i=1}^T\big). \tag{25}$$

The MAD provides a robust measure of score dispersion: a large MAD indicates that scores are spread over a wide range (clear bimodality), whereas a small MAD implies that scores cluster tightly (high ambiguity).

Step 2: Adaptive radius. We convert the MAD into a half-width radius r via

$$r = r_{\min} + (r_{\max} - r_{\min}) \cdot \sigma\big(-\lambda_r \cdot \big(\text{MAD} - \tau_r\big)\big), \tag{26}$$

where $\sigma(\cdot)$ is the logistic sigmoid, τ_r is a reference MAD threshold, $\lambda_r > 0$ controls the transition sharpness, and $[r_{\min}, r_{\max}]$ bounds the radius. The key behaviour is:

- When $\text{MAD} \gg \tau_r$ (bimodal distribution, clear separation), $\sigma(\cdot) \to 0$, so $r \to r_{\min}$: a *narrow* ambiguity interval, because few clips genuinely lie near the boundary.
- When $\text{MAD} \ll \tau_r$ (unimodal distribution, scores near 0.5), $\sigma(\cdot) \to 1$, so $r \to r_{\max}$: a *wide* ambiguity interval, capturing more clips for correction.

Step 3: Interval construction and clipping. The raw interval is centred at 0.5 (the natural decision boundary of the vMF score) and clipped to $[0, 1]$:

$$\rho_{\text{low}} = \max(0,\ 0.5 - r), \qquad \rho_{\text{high}} = \min(1,\ 0.5 + r). \tag{27}$$

Step 4: Tri-classification. Each clip is assigned to one of three categories based on its initial score:

$$c_i^{\text{tri}} = \begin{cases} \text{norm} & \text{if } s_i^{\text{init}} < \rho_{\text{low}}, \\ \text{amb} & \text{if } \rho_{\text{low}} \le s_i^{\text{init}} \le \rho_{\text{high}}, \\ \text{abn} & \text{if } s_i^{\text{init}} > \rho_{\text{high}}. \end{cases} \tag{28}$$

We denote the resulting clip sets as $\mathcal{C}_{\text{norm}}$, $\mathcal{C}_{\text{amb}}$, and $\mathcal{C}_{\text{abn}}$, respectively. Only clips in $\mathcal{C}_{\text{amb}}$ undergo geodesic pulling; clips in $\mathcal{C}_{\text{norm}}$ and $\mathcal{C}_{\text{abn}}$ retain their original features unchanged.

Practical defaults. In all experiments we use $r_{\min} = 0.05$, $r_{\max} = 0.25$, $\lambda_r = 20$, $\tau_r = 0.08$. These values were selected on a held-out subset of synthetic data and held fixed across all three benchmarks.

D.2 Dominant Prototype Voting

SGP's geodesic pulling requires a *target direction* for each ambiguous clip. Rather than pulling toward all prototypes simultaneously (which would average out discriminative directions), each video selects a small set of *dominant prototypes* that best represent the prevailing event types within that video. This design reflects the empirical observation that a given surveillance video typically contains at most one type of anomaly (e.g., a single fight or a single traffic accident), though normal behaviour may be diverse (e.g., walking, standing, vehicles passing).

Anomalous dominant prototype ($N_{\text{dom}}^{\text{abn}} = 1$). Each clip $i \in \mathcal{C}_{\text{abn}}$ votes for its nearest anomalous prototype:

$$k_i^{\text{abn}} = \underset{k \in \{1,\ldots,K_A\}}{\arg\max}\ \hat{f}_i^{l,\text{enh}\,\top} \boldsymbol{\mu}_k^{\text{abn}}. \tag{29}$$

The dominant anomalous prototype is determined by majority vote:

$$\boldsymbol{\mu}_{\text{dom}}^{\text{abn}} = \boldsymbol{\mu}_{k^*}^{\text{abn}}, \qquad k^* = \arg\max_k \left|\{i \in \mathcal{C}_{\text{abn}} : k_i^{\text{abn}} = k\}\right|. \tag{30}$$

Selecting a single dominant anomalous prototype enforces the prior that one anomaly type prevails per video, preventing ambiguous clips from being pulled toward unrelated anomaly directions.

Normal dominant prototypes ($N_{\text{dom}}^{\text{norm}} \le 2$). Normal behaviour is inherently more diverse, so we allow up to two dominant normal prototypes. Each clip $i \in \mathcal{C}_{\text{norm}}$ votes for its nearest normal prototype analogously to Eq. (29). The top-2 prototypes by vote count are retained:

$$\mathcal{I}_{\text{norm}} = \underset{k}{\text{TopK}}\left(\left|\{i \in \mathcal{C}_{\text{norm}} : k_i^{\text{norm}} = k\}\right|, 2\right), \qquad \mathcal{P}_{\text{dom}}^{\text{norm}} = \{\boldsymbol{\mu}_k^{\text{norm}} : k \in \mathcal{I}_{\text{norm}}\}. \tag{31}$$

If all normal clips vote for the same prototype, only one dominant normal prototype is used ($N_{\text{dom}}^{\text{norm}} = 1$).

Minimum anomalous clip threshold. If the number of anomalous clips is too small, the voting signal is unreliable and the video is treated as entirely normal:

$$|\mathcal{C}_{\text{abn}}| < n_{\min}^{\text{abn}} \implies \text{treat entire video as fully normal.} \tag{32}$$

When this condition is triggered, rather than skipping SGP, ambiguous clips are pulled exclusively toward the dominant normal prototype(s) to suppress spurious false positives, while anomalous prototypes are excluded from the pull targets. The threshold $n_{\min}^{\text{abn}}$ is set proportional to the total clip count: $n_{\min}^{\text{abn}} = \max\big(3,\ \lfloor \gamma_{\min} \cdot T \rfloor\big)$, where $\gamma_{\min} = 0.02$ is the minimum anomalous fraction. This prevents SGP from hallucinating anomalies in benign videos based on a handful of noisy high-scoring clips.

Fully normal video handling. If $\mathcal{C}_{\text{abn}} = \varnothing$ (no clip exceeds ρ_{high}), the video is automatically categorized as completely normal. In this case, all clips in $\mathcal{C}_{\text{amb}}$ are pulled exclusively toward the dominant normal prototype(s), suppressing residual score noise near the boundary.

D.3 Neighbour Consensus Assignment

Each ambiguous clip must be assigned a binary label $c_i \in \{\text{norm}, \text{abn}\}$ before pulling. Since the initial score s_i^{init} of an ambiguous clip is by definition unreliable (it lies in $[\rho_{\text{low}}, \rho_{\text{high}}]$), we resolve the ambiguity via *neighbour consensus*: the assignment is determined not by the clip's own feature but by the aggregated evidence from its visually similar neighbours within the same video.

Step 1: Intra-video sparse attention matrix. We construct a sparse attention matrix $\mathbf{A}_H \in \mathbb{R}^{T\times T}$ from the pairwise cosine similarities of the visual features $\{\hat{f}_i^v\}_{i=1}^T$ within the video:

$$\tilde{A}_{ij} = \hat{f}_i^{v\top} \hat{f}_j^v, \qquad i, j \in \{1, \dots, T\}. \tag{33}$$

The raw similarity matrix is sparsified by two mechanisms applied sequentially:

1. **Threshold sparsification.** Set $\tilde{A}_{ij} = 0$ if $\tilde{A}_{ij} < \tau_H$, where τ_H is a scene-similarity threshold. This removes connections between visually dissimilar clips (e.g., different camera angles or scene transitions within the same video).
2. **Top-K_H truncation.** For each row i, retain only the K_H largest entries and zero out the rest. This ensures that each clip attends to a bounded number of neighbours, preventing any single clip from being dominated by a large homogeneous group.

Self-connections are excluded ($\tilde{A}_{ii} = 0$) to prevent a clip from reinforcing its own potentially incorrect initial assessment. Finally, each row is normalised via softmax over the non-zero entries:

$$A_{ij}^H = \begin{cases} \dfrac{\exp(\tilde{A}_{ij}/\tau_{\text{temp}})}{\sum_{j':\tilde{A}_{ij'}>0} \exp(\tilde{A}_{ij'}/\tau_{\text{temp}})} & \text{if } \tilde{A}_{ij} > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

where $\tau_{\text{temp}} > 0$ is a temperature parameter controlling the attention sharpness.

Step 2: Aggregated neighbour feature. For each ambiguous clip $i \in \mathcal{C}_{\text{amb}}$, we compute the attention-weighted aggregation of its neighbours' *main features* (not visual features):

$$\hat{g}_i = \frac{\sum_{j=1}^T A_{ij}^H \, \hat{f}_j^{l,\text{enh}}}{\left\| \sum_{j=1}^T A_{ij}^H \, \hat{f}_j^{l,\text{enh}} \right\|_2} \in \mathcal{S}^{D-1}. \tag{35}$$

The ℓ_2-normalisation ensures that the aggregated feature remains on the sphere, enabling direct cosine comparison with the dominant prototypes.

Step 3: Prototype comparison and label assignment. The aggregated neighbour feature $\hat{g}_i$ is compared with the dominant prototypes by cosine similarity:

$$\text{sim}_{\text{norm}} = \max_{m\in\{1,\dots,N_{\text{dom}}^{\text{norm}}\}} \hat{g}_i^\top \boldsymbol{\mu}_{\text{dom},m}^{\text{norm}}, \qquad \text{sim}_{\text{abn}} = \hat{g}_i^\top \boldsymbol{\mu}_{\text{dom}}^{\text{abn}}. \tag{36}$$

The label is assigned as:

$$c_i = \begin{cases} \text{abn} & \text{if } \text{sim}_{\text{abn}} > \text{sim}_{\text{norm}} + \delta_{\text{margin}}, \\ \text{norm} & \text{otherwise,} \end{cases} \tag{37}$$

where $\delta_{\text{margin}} \geq 0$ is a small conservatism margin that biases toward the normal label in genuinely ambiguous cases, reflecting the prior that most clips in a surveillance video are normal. In practice, $\delta_{\text{margin}} = 0.01$.

Fallback for isolated clips. If an ambiguous clip i has no valid neighbours after sparsification (i.e., all entries in row i of $\mathbf{A}_H$ are zero), the neighbour consensus mechanism cannot be applied. In this case, the clip's *own* main feature $\hat{f}_i^{l,\text{enh}}$ is directly compared with the dominant prototypes:

$$c_i = \begin{cases} \text{abn} & \text{if } \hat{f}_i^{l,\text{enh}\top} \boldsymbol{\mu}_{\text{dom}}^{\text{abn}} > \max_m \hat{f}_i^{l,\text{enh}\top} \boldsymbol{\mu}_{\text{dom},m}^{\text{norm}} + \delta_{\text{margin}}, \\ \text{norm} & \text{otherwise.} \end{cases} \tag{38}$$

This fallback is conservative (biased toward normal) and occurs rarely in practice ($<2\%$ of ambiguous clips across all benchmarks).

Design rationale. The neighbour consensus mechanism rests on the assumption that *visually similar clips within the same video are likely to share the same anomaly status*. This is empirically well-grounded: surveillance videos exhibit strong temporal and visual continuity, so a clip depicting a fight scene is surrounded by other clips of the same fight, and a clip of normal pedestrian traffic is surrounded by similar normal clips. By aggregating evidence from neighbours (via their main features, which encode anomaly semantics) weighted by visual similarity (from visual features, which encode scene appearance), the assignment leverages complementary information from both feature streams and is robust to the noise in any single clip's initial score.

D.4 Score-Adaptive SLERP Derivation

Once each ambiguous clip $i \in \mathcal{C}_{\text{amb}}$ is assigned a label c_i and a corresponding target dominant prototype $\boldsymbol{\mu}_{\text{dom}}^{c_i}$, its feature is pulled along the geodesic (great-circle arc) toward the target via Spherical Linear Interpolation (SLERP) [39].

SLERP formula. For two unit vectors $\mathbf{p}, \mathbf{q} \in \mathcal{S}^{D-1}$ with geodesic angle $\Omega = \arccos(\mathbf{p}^\top \mathbf{q}) \in (0, \pi)$, the SLERP at parameter $t \in [0, 1]$ is defined as:

$$\text{Slerp}(\mathbf{p}, \mathbf{q}, t) = \frac{\sin\big((1-t)\,\Omega\big)}{\sin \Omega}\,\mathbf{p} + \frac{\sin(t\,\Omega)}{\sin \Omega}\,\mathbf{q}. \tag{39}$$

One can verify that $\|\text{Slerp}(\mathbf{p}, \mathbf{q}, t)\| = 1$ for all $t \in [0, 1]$ (the result always lies on $\mathcal{S}^{D-1}$), $\text{Slerp}(\mathbf{p}, \mathbf{q}, 0) = \mathbf{p}$, and $\text{Slerp}(\mathbf{p}, \mathbf{q}, 1) = \mathbf{q}$. Furthermore, the angular velocity along the geodesic is constant: $d_{\text{geo}}\big(\mathbf{p}, \text{Slerp}(\mathbf{p}, \mathbf{q}, t)\big) = t\,\Omega$. When $\Omega \to 0$ (nearly identical vectors), SLERP degenerates to LERP with normalisation; our implementation handles this edge case by switching to ℓ_2-normalised LERP when $\sin \Omega < 10^{-6}$.

Comparison with Normalized LERP. The naive Euclidean linear interpolation (LERP) followed by ℓ_2-normalization (NLERP) is defined as:

$$\text{NLerp}(p, q, t) = \frac{(1-t)p + tq}{\|(1-t)p + tq\|_2}. \tag{40}$$

While $\text{NLerp}(p, q, 0.5)$ coincides with the geodesic midpoint $\text{Slerp}(p, q, 0.5)$, NLERP fails to traverse the great-circle arc at constant angular velocity for arbitrary $t \in (0, 1)$. Specifically, the angular distance traversed by NLERP from p satisfies:

$$\cos\theta(t) = \langle p, \text{NLerp}(p, q, t)\rangle = \frac{1 - t + t\cos\Omega}{\sqrt{(1-t)^2 + t^2 + 2t(1-t)\cos\Omega}}. \tag{41}$$

The discrepancy between the constant-speed geodesic arc $\text{Slerp}(p, q, t)$ and NLERP is given by $\Delta\theta(t) = |t\Omega - \theta(t)|$. Expanding $\Delta\theta(t)$ around $t = 0.5$ reveals an intrinsic angular acceleration:

$$\Delta\theta(t) = \frac{\Omega^3}{12}\left|t - \frac{1}{2}\right|\left(t - t^2\right) + \mathcal{O}(\Omega^5). \tag{42}$$

Crucially, NLERP projects Euclidean secant chords back onto the hypersphere, causing uneven angular progression where features accelerate near midpoints and decelerate near endpoints. SLERP eliminates this chordal distortion by construction, strictly preserving constant geodesic velocity and manifold constraints on $\mathcal{S}^{D-1}$.

Score-adaptive pull strength. A uniform pull strength β for all ambiguous clips would be suboptimal: a clip with $s_i^{\text{init}} = 0.50$ (maximally uncertain) should receive a stronger correction than a clip with $s_i^{\text{init}} = 0.42$ (mildly uncertain, leaning normal). We define a normalised distance-from-boundary measure:

$$\hat{d}_i = \begin{cases} \dfrac{s_i^{\text{init}} - 0.5}{\rho_{\text{high}} - 0.5} & \text{if } s_i^{\text{init}} \geq 0.5, \\ \dfrac{0.5 - s_i^{\text{init}}}{0.5 - \rho_{\text{low}}} & \text{if } s_i^{\text{init}} < 0.5. \end{cases} \tag{43}$$

By construction, $\hat{d}_i \in [0, 1]$: $\hat{d}_i = 0$ when $s_i^{\text{init}} = 0.5$ (maximum ambiguity) and $\hat{d}_i = 1$ when $s_i^{\text{init}} \in \{\rho_{\text{low}}, \rho_{\text{high}}\}$ (boundary of the ambiguity interval, minimum ambiguity). The SLERP interpolation parameter for clip i is then:

$$\beta_i = \beta_{\text{base}} \times \big(1 - \tfrac{1}{2}\hat{d}_i\big), \tag{44}$$

where $\beta_{\text{base}} \in (0, 1)$ is a global hyperparameter controlling the maximum pull strength. The factor $(1 - \frac{1}{2}\hat{d}_i)$ ensures:

- $\hat{d}_i = 0$ (most ambiguous, $s \approx 0.5$): $\beta_i = \beta_{\text{base}}$ (full pull strength).
- $\hat{d}_i = 1$ (least ambiguous, $s \approx \rho_{\text{low}}$ or ρ_{high}): $\beta_i = \frac{1}{2}\beta_{\text{base}}$ (half pull strength).

The pull strength thus decreases linearly with decreasing uncertainty, preventing over-correction of clips whose initial scores already lean toward the correct label. The factor $\frac{1}{2}$ prevents the pull from vanishing entirely at the interval boundaries, ensuring that even mildly ambiguous clips receive a non-trivial correction.

Feature pulling. The pulled feature for each ambiguous clip is computed as:

$$\hat{f}_i^{\text{pulled}} = \text{Slerp}\big(\hat{f}_i^{l,\text{enh}},\ \boldsymbol{\mu}_{\text{target}}^{c_i},\ \beta_i\big), \tag{45}$$

where $\boldsymbol{\mu}_{\text{target}}^{c_i}$ is the dominant prototype corresponding to the assigned label c_i. For clips assigned $c_i = \text{abn}$, the target is $\boldsymbol{\mu}_{\text{target}}^{\text{abn}} = \boldsymbol{\mu}_{\text{dom}}^{\text{abn}}$. For clips assigned $c_i = \text{norm}$, the target is the nearest dominant normal prototype:

$$m_i^* = \underset{m \in \{1,\ldots,N_{\text{dom}}^{\text{norm}}\}}{\arg\max} \big(\hat{f}_i^{l,\text{enh}}\big)^\top \boldsymbol{\mu}_{\text{dom},m}^{\text{norm}}, \qquad \boldsymbol{\mu}_{\text{target}}^{\text{norm}} = \boldsymbol{\mu}_{\text{dom},m_i^*}^{\text{norm}}. \tag{46}$$

The pulled feature lies on $\mathcal{S}^{D-1}$ by the SLERP guarantee, preserving the manifold constraint. Clips outside $\mathcal{C}_{\text{amb}}$ retain their enhanced features: $\hat{f}_i^{\text{pulled}} = \hat{f}_i^{l,\text{enh}}$ for $i \notin \mathcal{C}_{\text{amb}}$.

Geometric interpretation. Figure 4(c–d) in the main text visualises the effect of SGP on the sphere. Before pulling, ambiguous features scatter between the normal and anomalous prototype clusters, often with incorrect implicit labels. After pulling, each ambiguous feature migrates along the geodesic toward its assigned prototype by an arc length of $\beta_i \cdot \Omega_i$, where $\Omega_i = d_{\text{geo}}(\hat{f}_i^{l,\text{enh}}, \boldsymbol{\mu}_{\text{target}}^{c_i})$. This produces a clearly bimodal feature distribution on the sphere, which translates directly into a bimodal score distribution with reduced overlap—as confirmed both quantitatively (Table 2, M2→M3) and qualitatively (score curves in Figure 4(a–b)).

Why only ambiguous clips are pulled. Pulling clips with confident classifications ($s \ll \rho_{\text{low}}$ or $s \gg \rho_{\text{high}}$) is both unnecessary and potentially harmful. These clips already have strong initial scores, and pulling them toward prototypes could *over-regularise* their features, erasing fine-grained distinctions (e.g., between different intensities of the same anomaly type). The tri-classification mechanism ensures that SGP intervenes only where it is needed—on the decision boundary—while leaving well-separated clips undisturbed.

D.5 Complete Algorithm

Algorithm 1 presents the complete SGP procedure for a single video, corresponding to pipeline Stage (d) in Figure 2.

Algorithm 1 vMF-Guided Spherical Geodesic Pulling (SGP) — Single Video

Require: $\{\hat{f}_i^{l,\text{enh}}\}_{i=1}^T$, $\{\hat{f}_i^v\}_{i=1}^T \subset \mathcal{S}^{D-1}$; scores $\{s_i^{\text{init}}\} \subset [0,1]$; prototypes $\{\boldsymbol{\mu}_k^{\text{norm}}\}_{k=1}^{K_N}$, $\{\boldsymbol{\mu}_k^{\text{abn}}\}_{k=1}^{K_A}$; params $\beta_{\text{base}}, r_{\min}, r_{\max}, \lambda_r, \tau_r, \gamma_{\min}, \delta_{\text{margin}}, \tau_H, K_H, \tau_{\text{temp}}$
Ensure: $\{s_i^{\text{final}}\}_{i=1}^T \subset [0,1]$

— Stage 1: MAD-Based Adaptive Tri-Classification —
$\tilde{s} \leftarrow \text{median}(\{s_i^{\text{init}}\})$; $\text{MAD} \leftarrow \text{median}(\{|s_i^{\text{init}} - \tilde{s}|\})$
$r \leftarrow r_{\min} + (r_{\max} - r_{\min}) \cdot \sigma(-\lambda_r(\text{MAD} - \tau_r))$; $\rho_{\text{lo}} \leftarrow \max(0, 0.5 - r)$; $\rho_{\text{hi}} \leftarrow \min(1, 0.5 + r)$
$\mathcal{C}_{\text{N}} \leftarrow \{i : s_i < \rho_{\text{lo}}\}$; $\mathcal{C}_{\text{A}} \leftarrow \{i : s_i > \rho_{\text{hi}}\}$; $\mathcal{C}_{\text{amb}} \leftarrow \{1, \ldots, T\} \setminus (\mathcal{C}_{\text{N}} \cup \mathcal{C}_{\text{A}})$
— Stage 2: Dominant Prototype Voting —
$n_{\min}^{\text{abn}} \leftarrow \max(3, \lfloor \gamma_{\min} T \rfloor)$
if $|\mathcal{C}_{\text{A}}| < n_{\min}^{\text{abn}}$ **then**
    Treat video as fully normal; pull all $\mathcal{C}_{\text{amb}}$ toward dominant normal prototype(s), score via Eq. (4), **return**
end if
Each $i \in \mathcal{C}_{\text{A}}$ votes $k_i^{\text{abn}} = \arg\max_k \hat{f}_i^{l,\text{enh}\top} \boldsymbol{\mu}_k^{\text{abn}}$; $\boldsymbol{\mu}_{\text{dom}}^{\text{abn}} \leftarrow$ majority winner
Each $i \in \mathcal{C}_{\text{N}}$ votes similarly; retain top-2 normal prototypes $\boldsymbol{\mu}_{\text{dom},1}^{\text{norm}}, \boldsymbol{\mu}_{\text{dom},2}^{\text{norm}}$
— Stage 3: Neighbour Consensus Assignment —
Build sparse intra-video attention $\mathbf{A}_H$ from $\{\hat{f}_i^v\}$ (threshold τ_H, top-K_H, softmax)
for each $i \in \mathcal{C}_{\text{amb}}$ **do**
    if $\sum_j A_{ij}^H > 0$ **then**
        $\hat{g}_i \leftarrow (\sum_j A_{ij}^H \hat{f}_j^{l,\text{enh}}) / \| \cdot \|_2$ {neighbour aggregate}
    else
        $\hat{g}_i \leftarrow \hat{f}_i^{l,\text{enh}}$ {self-fallback}
    end if
    $c_i \leftarrow \text{abn}$ if $\hat{g}_i^\top \boldsymbol{\mu}_{\text{dom}}^{\text{abn}} > \max_m \hat{g}_i^\top \boldsymbol{\mu}_{\text{dom},m}^{\text{norm}} + \delta_{\text{margin}}$, else $c_i \leftarrow \text{norm}$
end for
— Stage 4: Score-Adaptive SLERP Pulling —
for each $i \in \mathcal{C}_{\text{amb}}$ **do**
    $\hat{d}_i$ via Eq. (43); $\beta_i \leftarrow \beta_{\text{base}}(1 - \frac{1}{2}\hat{d}_i)$; $\Omega_i \leftarrow \arccos(\hat{f}_i^{l,\text{enh}\top} \boldsymbol{\mu}_{\text{target}}^{c_i})$
    $\hat{f}_i^{\text{pulled}} \leftarrow \text{Slerp}(\hat{f}_i^{l,\text{enh}}, \boldsymbol{\mu}_{\text{target}}^{c_i}, \beta_i)$ {Eq. (39); skip if $\sin \Omega_i < 10^{-6}$}
end for
Non-ambiguous clips: $\hat{f}_i^{\text{pulled}} \leftarrow \hat{f}_i^{l,\text{enh}}$ for $i \notin \mathcal{C}_{\text{amb}}$
— Final Scoring —
for each $i \in \{1, \ldots, T\}$ **do**
    $d_{\text{norm}} = \min_k \arccos(\hat{f}_i^{\text{pulled}\top} \boldsymbol{\mu}_k^{\text{norm}})$; $d_{\text{abn}} = \min_k \arccos(\hat{f}_i^{\text{pulled}\top} \boldsymbol{\mu}_k^{\text{abn}})$
    $s_i^{\text{final}} \leftarrow \sigma(\kappa(d_{\text{norm}} - d_{\text{abn}}))$ {Eq. (4)}
end for
return $\{s_i^{\text{final}}\}_{i=1}^T$

Computational complexity. For a video with T clips and feature dimension D, the dominant costs are: (i) the intra-video attention matrix $\mathbf{A}_H$: $O(T^2 D)$ for pairwise cosine similarities, reduced to $O(T \cdot K_H \cdot D)$ after top-K_H sparsification; (ii) neighbour aggregation: $O(|\mathcal{C}_{\text{amb}}| \cdot K_H \cdot D)$; (iii) SLERP computation: $O(|\mathcal{C}_{\text{amb}}| \cdot D)$. In practice, $|\mathcal{C}_{\text{amb}}| \ll T$ (typically10–25% of clips), $K_H \leq 20$, and $D = 4096$, so the total SGP cost per video is negligible compared to the MLLM forward pass. For the largest videos in XD-Violence ($T \approx 1{,}800$ clips), SGP completes in $<$0.3 seconds on a single CPU core.

UBnormal exception. As noted in Table 2, SGP is not applied to UBnormal. The 211 test videos in UBnormal average only $\sim$12 seconds each ($\sim$10–14 clips per video), providing insufficient temporal context for the intra-video tri-classification and neighbour consensus mechanisms that SGP relies on. With so few clips per video, the MAD-based ambiguity interval becomes unreliable (the median and MAD of 5–8 values are highly variable), and the intra-video attention matrix degenerates to a near-complete graph where neighbour consensus offers no advantage over self-assignment. Consequently, M2 and M3 produce identical results on UBnormal.

E Computational Cost Analysis

A key advantage of SphereVAD over existing training-free VAD methods is that it requires *no LLM or VLM inference calls at test time*—all computation operates on pre-extracted intermediate-layer features via lightweight geometric operations. This section provides a comprehensive wall-clock time

and storage analysis, covering both **offline** (batch processing) and **online** (streaming) deployment modes.

E.1 Offline Inference

In the offline setting, SphereVAD processes an entire test set in two sequential phases: (1) a one-time feature extraction pass through the frozen MLLM, and (2) the geometric inference pipeline (Stages S1–S6). Both phases are fully parallelisable and require no gradient computation.

E.1.1 Feature Extraction (One-Time Cost)

Table 9 reports the wall-clock time for extracting dual-stream features ($\hat{f}^l$ and $\hat{f}^v$) from the frozen Qwen3.5 backbone across all datasets, including the synthetic calibration set. Each clip produces two ℓ_2-normalised vectors of dimension $D{=}4096$.

Table 9: Feature extraction time and storage cost. Times are reported for single-GPU (1×A100-80GB) and multi-GPU (4×A100-80GB) configurations. Storage is computed as Clips $\times$ 2 $\times$ D $\times$ 4 bytes (float32) using standard decimal GB.

Dataset	**Clips**	**1×A100**	**4×A100**	**Storage**
SYN (calibration)	2,040	5.5 min	1.5 min	0.07 GB
UCF-Crime	46,410	2.1 h	33.8 min	1.52 GB
XD-Violence	97,396	4.3 h	70.9 min	3.19 GB
UBnormal	3,912	10.5 min	2.9 min	0.13 GB
Total	149,758	∼6.7 h	∼1.8 h (109.1 min)	4.91 GB

The per-clip storage cost is:

$$2 \text{ (streams)} \times 4096 \text{ (dim)} \times 4 \text{ (bytes/float32)} = 32{,}768 \text{ bytes} \approx 32 \text{ KB/clip}. \tag{47}$$

Even for the largest benchmark (XD-Violence, ∼97K clips), the total feature storage is only ∼3.19 GB, easily fitting in CPU memory. With 4×A100 GPUs and asynchronous prefetching (Appendix C.4), the complete extraction for all benchmarks plus calibration data finishes in under 2 hours—a *one-time* cost that is amortised over all subsequent experiments, hyperparameter sweeps, and ablation studies.

E.1.2 Geometric Inference Pipeline

Once features are extracted, the six-stage geometric inference pipeline (Stages S1–S6; cf. Figure 2) operates entirely on CPU with no GPU requirement. Table 10 provides a per-stage wall-clock breakdown measured on a single CPU core (Intel Xeon Platinum 8375C).

Table 10: Per-stage wall-clock time (seconds) for the geometric inference pipeline. All stages run on a single CPU core after features have been loaded into memory. Data loading time is excluded.

Stage	**XD-Violence**	**UCF-Crime**	**UBnormal**
S1: Fréchet Mean + Spherical Centering	0.591 s	0.219 s	0.226 s
S2: vMF Prototype Construction (K-Means)	0.480 s	0.362 s	0.388 s
S3: HSA (Cross-Video Holistic Attention)	9.218 s	2.088 s	0.047 s
S4: Initial vMF Scoring	0.119 s	0.040 s	0.028 s
S5: SGP (Spherical Geodesic Pulling)	2.528 s	0.735 s	—[†]
S6: Final Scoring + Gaussian Smoothing	2.957 s	1.032 s	0.288 s
Total Inference	**15.893 s**	**4.476 s**	**0.977 s**
Data Loading (excluded)	5.08 s	3.50 s	2.00 s

[†] SGP is disabled on UBnormal due to short average video duration.

Key observations.

- The *entire* geometric inference pipeline completes in **under 16 seconds** on the largest benchmark (XD-Violence, ∼97K clips), and under **5 seconds** on UCF-Crime (∼46K clips). This is orders of magnitude faster than any LLM/VLM-based scoring method.
- The dominant cost is **HSA** (Stage S3), which constructs the cross-video sparse attention matrix. Its $O(N_{\text{total}}^2 \cdot D)$ pairwise cosine similarity computation scales quadratically with the number of test clips, accounting for ∼58% of the total time on XD-Violence. For UBnormal (∼3.9K clips), HSA is negligible (0.047 s).
- **SGP** (Stage S5) is applied per-video and scales linearly with the number of clips per video; its total time across all test videos is modest (<3 s even on XD-Violence).
- All stages are *deterministic* (no stochastic sampling) and require *no GPU*: the pipeline runs on a single CPU core, making it deployable in resource-constrained settings.

E.1.3 Comparison with Existing Training-Free Methods

Table 11 compares the computational profile of SphereVAD with three representative training-free VAD methods across multiple efficiency dimensions. All timings are normalised to the UCF-Crime test set for fair comparison.

Table 11: Computational cost comparison among training-free VAD methods on the UCF-Crime test set. "LLM/VLM Calls" counts the number of large-model inference queries required per test video. SphereVAD's inference pipeline requires *zero* such calls, operating entirely on pre-extracted features.

Method	LLM/VLM Calls (per video)	GPU Inference (total GPU-hours)	Post-Extraction Time	AUC (%)
PANDA [49]	Multiple (VLM + MLLM, iterative reflection) *Average speed: 0.82 FPS on A6000*	N/A†	—	84.89
VADTree [26]	VLM caption + LLM scoring per node	62.3 h (2×3090)	—	84.74
Unified-VAD [27]	1 VLM + 1–2 LLM per 16 frames *Amortised: 0.029 s/frame on 2×3090*	N/A†	—	84.28
SphereVAD (Ours)	**0** (feature extraction only)	2.25 GPU-h (33.8 min on 4×A100)‡	**4.5 s (CPU)**	**86.38**

†Exact total GPU-hours not reported; per-frame/FPS timings are cited from the original papers.
‡Feature extraction on UCF-Crime is a one-time cost (33.8 min ≈ 0.563 h × 4 GPUs ≈ 2.25 GPU-hours); subsequent experiments (ablations, hyperparameter sweeps) reuse cached features with zero GPU cost.

Architectural efficiency advantage. The fundamental computational distinction between SphereVAD and all prior training-free methods lies in the *absence of test-time LLM/VLM inference calls*. Methods such as PANDA [49], VADTree [26], and Unified-VAD [27] invoke large foundation models (VLMs for captioning/perception, LLMs for scoring/reasoning) repeatedly during inference—often multiple times per video clip—incurring substantial GPU cost that scales linearly (or worse) with the number of test clips. PANDA further compounds this with iterative reflection loops that invoke external tools, reducing its throughput to <1 FPS. VADTree requires over 62 GPU-hours for the VLM/LLM pipeline alone on UCF-Crime.

In contrast, SphereVAD performs a *single* forward pass through the frozen MLLM to extract intermediate-layer features (a one-time cost of 33.8 min on 4×A100 for UCF-Crime), after which *all subsequent computation*—spherical centering, prototype construction, HSA, vMF scoring, and SGP—is performed via closed-form geometric operations on pre-extracted features, completing in <5 seconds on a single CPU core. This separation of feature extraction from geometric inference yields two practical benefits: (i) the one-time extraction cost is amortised across all downstream experiments, enabling rapid iteration during development; and (ii) the inference pipeline itself is lightweight enough to run on commodity hardware without GPU access.

E.2 Online Inference

For real-time deployment scenarios (e.g., live surveillance feeds), SphereVAD supports a streamlined **online inference mode** that trades a small amount of detection accuracy for substantial latency reduction. The online pipeline simplifies the offline framework as follows:

1. **Calibration from synthetic data only.** The Fréchet mean μ_{unified} is computed using *only* the synthetic calibration features $\{\tilde{f}^l_{\text{syn}}\}$, without pooling real test-set features. This eliminates the need to accumulate statistics over the test set before inference begins—the spherical reference frame is fixed *a priori*, ensuring a strictly inductive, real-time setting.
2. **Prototypes from synthetic data only.** The vMF prototypes $\{\mu_k^{\text{norm}}\}$, $\{\mu_k^{\text{abn}}\}$ are constructed via spherical K-Means on the centered synthetic features, identical to the offline pipeline. No test-set statistics are required.
3. **Direct vMF scoring (no HSA, no SGP).** Each incoming clip is independently scored via the vMF likelihood-ratio criterion (Eq. (4)) immediately after feature extraction and spherical centering. The cross-video HSA module (which requires access to other test clips) and the intra-video SGP module (which requires a complete video sequence) are both disabled, as they are inherently batch operations.

Note that the online mode is a synthetic-only variant that differs from configuration **M1** in Table 2 of the main text: M1 uses the unified Fréchet mean computed from *both* synthetic and real features (Eq. (1)), whereas the online mode computes the Fréchet mean from synthetic data *only*. The online mode therefore constitutes a strictly more constrained, single-clip setting than M1.

Throughput analysis. Under this configuration, per-clip inference consists of three operations: (i) a single MLLM forward pass for dual-stream feature extraction, (ii) logarithmic-map centering and ℓ_2-normalisation (Eq. (2)), and (iii) vMF scoring via Eq. (4). Operations (ii) and (iii) take negligible runtime ($<$0.1 ms per clip). The throughput is therefore dominated by the forward pass.

On a single A100-80GB GPU with the Qwen3.5 backbone (processing an isolated single clip formatted as a 2×2 composite frame), the forward pass takes approximately $\sim$67.5 ms per clip without disk write or offline feature dumping. Assuming a standard surveillance video frame rate of 24 FPS and an input clip sampling of 4 frames:

$$\text{FPS}_{\text{online}} = \frac{4 \text{ frames/clip}}{0.0675 \text{ s/clip}} \approx \textbf{59.3 FPS}, \tag{48}$$

which exceeds real-time requirements (24 FPS) by a factor of $\sim$2.5$\times$, leaving sufficient headroom for video decoding and streaming I/O.

Online vs. offline accuracy. The online mode yields 81.37% AP on XD-Violence, 78.31% AUC on UCF-Crime, and 71.62% AUC on UBnormal. The accuracy gap relative to the full offline pipeline (M3: 86.99% AP, 86.38% AUC, 76.46% AUC) reflects the combined effect of: (i) relying solely on synthetic calibration without target-domain centering, which leaves a residual domain gap unabsorbed, and (ii) disabling HSA and SGP, which provide cross-video consistency and intra-video boundary refinement, respectively. For reference, configuration M1 in Table 2—which retains the unified (synthetic + real) Fréchet mean but likewise disables HSA and SGP—achieves 82.51% AP on XD-Violence, 81.68% AUC on UCF-Crime, and 75.98% AUC on UBnormal, illustrating the expected cost of synthetic-only inductive centering.

Nonetheless, on XD-Violence, this lightweight online configuration substantially outperforms prior training-free zero-shot baselines (e.g., LAVAD at 62.01% AP and EventVAD at 64.04% AP by over $+17$ pp). On UCF-Crime, while lower than offline multi-prompt architectures like LAVAD (80.28%), it maintains competitive streaming utility at > 59 FPS, contrasting sharply with autoregressive or agent-based video LLM approaches (e.g., PANDA) that typically operate below 1 FPS.

Summary. Table 12 contrasts the offline and online deployment modes.

F Additional Visualizations

F.1 Additional Frame-Level Score Curves

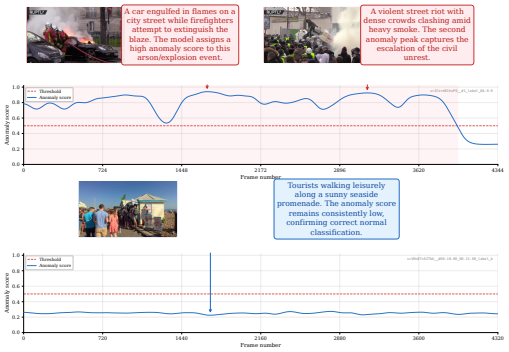
To complement the examples in the main text (Figure 4), we present four additional pairs of abnormal and normal videos from the XD-Violence and UCF-Crime datasets in Figure 8. Each pair consists of an upper panel showing an abnormal video with ground-truth annotations (shaded regions) and a lower panel showing a normal video. Representative frames are displayed alongside their textual

Table 12: Offline vs. online deployment modes of SphereVAD.

Mode	Fréchet Mean	Prototypes	HSA	SGP	Throughput	XD-Vio / UCF / UBn
Offline (M3)	Syn + Real	Syn	✓	✓*	4.5 s total (CPU)	86.99 / 86.38 / 76.46
Offline (M1)	Syn + Real	Syn	✗	✗	4.5 s total (CPU)	82.51 / 81.68 / 75.98
Online	Syn only	Syn	✗	✗	59.3 FPS (1×A100)	81.37 / 78.31 / 71.62

*SGP is disabled for UBnormal due to short video duration (see Appendix D).
Metrics: XD-Violence = AP(%); UCF-Crime and UBnormal = AUC(%).
M1 uses the unified (Syn + Real) Fréchet mean but disables HSA and SGP; the online mode further restricts the Fréchet mean to synthetic data only, enabling single-clip streaming without any test-set statistics.

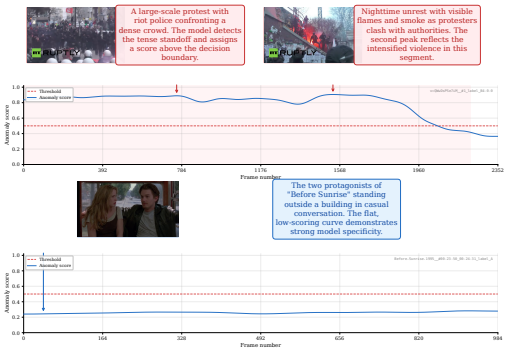
descriptions. The anomaly score curves demonstrate that our method consistently assigns high scores within annotated violent segments while maintaining low scores for normal content.



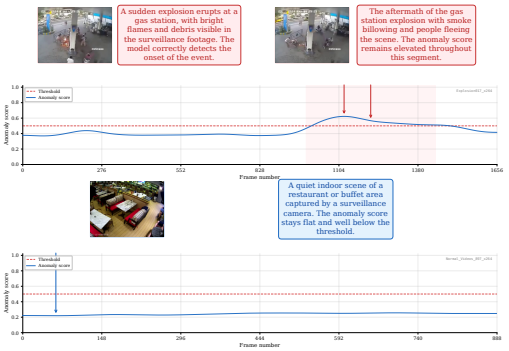
(a) *Top:* Car fire and street riot with two distinct score peaks. *Bottom:* Tourists at a seaside promenade (normal).



(b) *Top:* Street riot with pedestrians fleeing in panic. *Bottom:* Indoor dialogue from *Before Sunrise* (normal).



(c) *Top:* Large-scale protest escalating into nighttime unrest. *Bottom:* Outdoor conversation from *Before Sunrise* (normal).



(d) *Top:* Gas station explosion with flames and fleeing pedestrians. *Bottom:* Quiet indoor surveillance of a restaurant area (normal).

Figure 8: **Additional frame-level anomaly score curves** on XD-Violence (a–c) and UCF-Crime (d). Each sub-figure shows an abnormal video (upper panel, with GT shaded) paired with a normal video (lower panel). The model produces high scores aligned with annotated violent segments while maintaining low scores for normal content.

F.2 Additional SLERP Sphere Visualizations

To provide a comprehensive view of the SLERP geodesic pulling effect on the hypersphere, we render the same set of ambiguous clips from three different viewpoints in Figure 9. In each sub-figure, the left panel shows the ambiguous clip features *before* SLERP pulling (i.e., at the M2 stage), while the right panel shows the same features *after* being pulled toward their nearest vMF prototypes (M3

stage). All panels share the same PCA projection and color scale, where blue indicates low anomaly scores (normal) and red indicates high anomaly scores (abnormal). The visualizations confirm that SLERP consistently resolves ambiguous features by displacing them toward the correct prototype region, resulting in a clearer separation between normal and abnormal clusters across all viewing angles.

F.3 Domain Shift Visualization

Figure 10 visualises the distributional gap between the synthetic calibration domain and the three real test domains in both Euclidean and hyperspherical feature spaces, providing direct geometric evidence for the rotational bias discussed in Section 3.2 and the necessity of unified Fréchet mean centering.

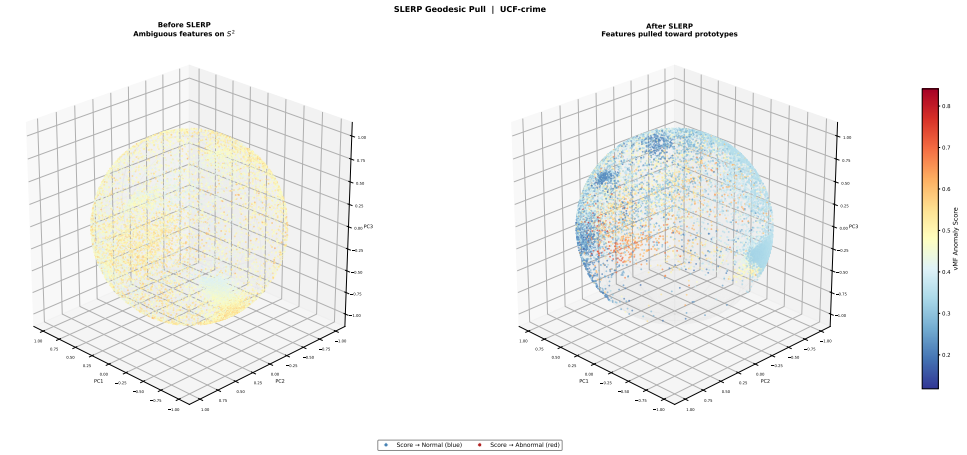
Setup. We extract main features f^l (last-token hidden states at layer $\ell^*{=}31$) from the frozen Qwen3.5 backbone for all clips in the synthetic calibration set and the three test sets (UCF-Crime, XD-Violence, UBnormal). For the Euclidean view (a), we compute the per-dataset arithmetic mean in the original high-dimensional space and project all means into 3D via PCA (fitted on a pooled subsample of $5{,}000$ clips per dataset). For the hyperspherical view (b), we ℓ_2-normalise all features onto $\mathcal{S}^{D-1}$, compute the per-dataset Fréchet mean via Karcher iteration (Appendix A.1), and project the four Fréchet means into 3D via a separate PCA fitted on the normalised features. To aid visual clarity, the Synthetic mean is placed at the origin (Euclidean) or the north pole (hypersphere), and the remaining three datasets are arranged at uniform azimuthal intervals (120° apart) while preserving the true inter-domain distances or geodesic angles, respectively.

Euclidean space (panel a). Each dataset is represented by a translucent sphere of uniform radius (the 50th-percentile coverage radius of UBnormal, used as a common scale reference) centred at its Euclidean mean (star marker). The Euclidean distances from Synthetic to the three test domains are $d_{\mathrm{UCF}}{=}21.74$, $d_{\mathrm{XD}}{=}14.29$, and $d_{\mathrm{UBnormal}}{=}17.34$. These distances reflect both directional and magnitude differences in the raw feature space and are not directly comparable across domains because the Euclidean metric conflates norm variation with angular displacement.

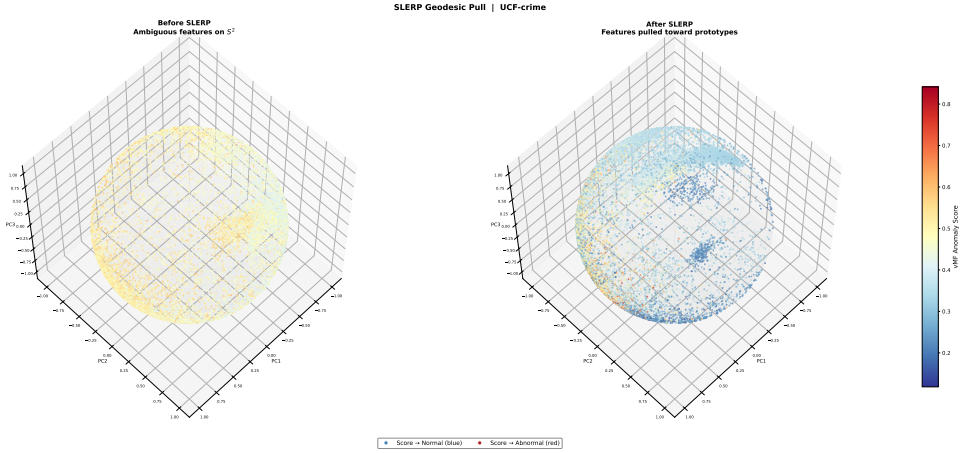
Hypersphere space (panel b). After ℓ_2-normalisation, all features reside on the unit hypersphere $\mathcal{S}^{D-1}$, and the natural metric becomes the geodesic (great-circle) angle. The geodesic gaps from Synthetic to the three test domains are $\Delta_{\mathrm{UCF}}{=}8.2^\circ$, $\Delta_{\mathrm{XD}}{=}5.4^\circ$, and $\Delta_{\mathrm{UBnormal}}{=}6.5^\circ$, visualised as arcs on the unit sphere connecting the Fréchet means. These offsets of 5°–8° confirm that the domain shift is predominantly *directional* rather than magnitude-based: once norms are factored out, the residual discrepancy reduces to a modest angular rotation.

Implications for spherical centering. The visualisation provides three key insights that motivate the unified Fréchet mean centering of Eq. (2):

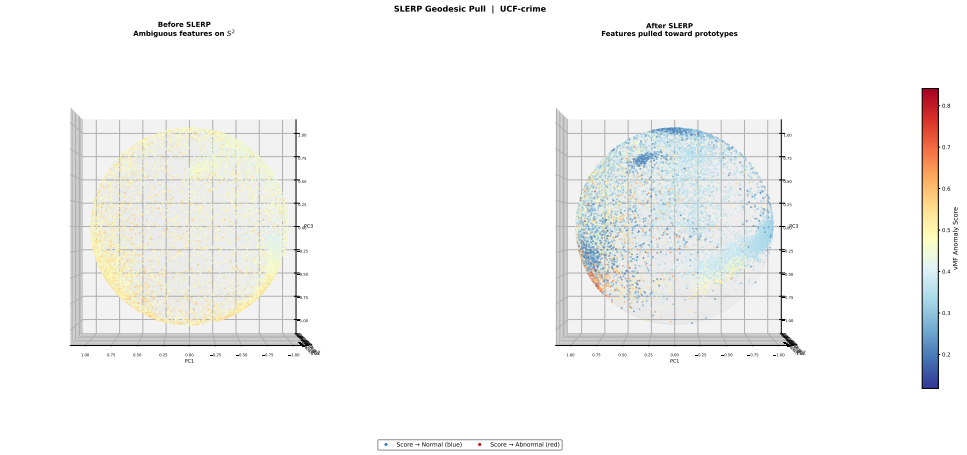
1. **The domain shift is geometrically compact on the sphere.** While Euclidean distances span a wide range (14–22), the geodesic angles cluster within 5°–8°, indicating that the four domain distributions occupy nearby but systematically offset spherical caps. A single centering operation can absorb this offset.
2. **Synthetic–real misalignment is universal.** All three test domains exhibit a non-negligible angular displacement from the synthetic calibration domain. Without centering, vMF prototypes calibrated on synthetic features would be systematically rotated away from the real feature distribution, degrading scoring accuracy—consistent with the $+5.78\,\%$ AUC gain observed on UCF-Crime upon adding spherical centering (Table 2, M0a$\rightarrow$M0b).
3. **The unified Fréchet mean provides a symmetric reference.** By pooling synthetic and real features before computing the Fréchet mean (Eq. (1)), the centering base point lies near the geodesic midpoint of the domain-specific means (Proposition 2), absorbing the rotational bias symmetrically from both sides rather than privileging either domain.



(a) Default view (elev=25°, azim=135°). The overall distribution of ambiguous clips is visible, with SLERP pulling features away from the decision boundary toward their respective prototype clusters.



(b) Top-down view (elev=60°, azim=45°). This bird's-eye perspective highlights the lateral displacement of ambiguous features, showing that clips initially clustered near the equatorial boundary are redistributed toward polar regions after SLERP.



(c) Side view (elev=0°, azim=90°). The profile view reveals the geodesic displacement along the vertical axis, confirming that the pulling magnitude is larger for clips closer to the decision boundary (score ≈ 0.5).

Figure 9: **SLERP geodesic pulling visualized from three viewpoints** on UCF-Crime. *Left* of each panel: ambiguous features before SLERP (M2 stage). *Right* of each panel: the same features after SLERP pulling toward vMF prototypes (M3 stage). Points are colored by their vMF anomaly score (blue = normal, red = abnormal). The top 50% of ambiguous clips ranked by geodesic displacement are shown.

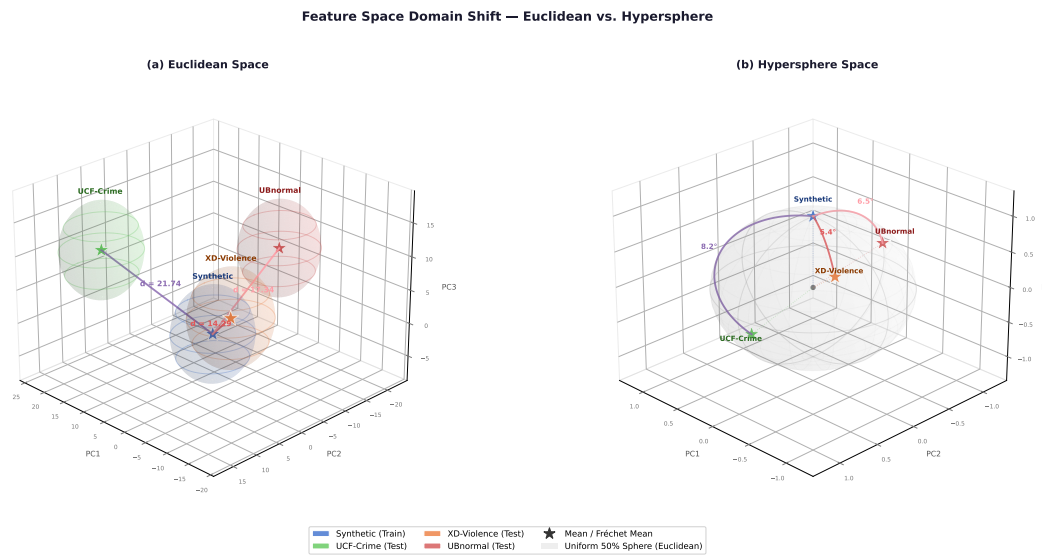


Figure 10: **Domain shift between synthetic and real feature distributions.** (a) *Euclidean space*: per-dataset arithmetic means (stars) with uniform 50%-coverage spheres; straight lines show Euclidean distances from Synthetic to each test domain. (b) *Hypersphere space*: per-dataset Fréchet means (stars) on the unit sphere $\mathcal{S}^{D-1}$; geodesic arcs show angular offsets of $5°$–$8°$, confirming a systematic rotational bias that is absorbed by unified Fréchet mean centering (Section 3.2). Features are extracted from layer 31 of a frozen Qwen3.5; 3D positions are obtained via PCA and rearranged for visual clarity (Synthetic at origin/north pole, others at $120°$ azimuthal intervals) while preserving true inter-domain distances/angles.

G Hyperparameter Sensitivity Analysis

SphereVAD contains four dataset-specific core hyperparameters: the normal and anomalous prototype counts (K_N, K_A), the HSA mixing coefficient α_G, and the SLERP base pull strength β_{base}. Note that β_{base} is applicable only to XD-Violence and UCF-Crime, since SGP is disabled for UBnormal (whose test videos are too short for reliable intra-video tri-classification; see Appendix D and Table 2 in the main text). The following subsections list the configurations evaluated across all three benchmarks, provide detailed sensitivity sweeps for each component, and report a unified-setting evaluation demonstrating the robustness of the framework.

G.1 Complete Hyperparameter Table

Table 13 summarises the hyperparameter configurations evaluated across the three benchmark datasets. All other algorithmic parameters (MAD scale factors, attention sparsification constants, etc.) are fixed globally across all datasets; their specific values and descriptions are detailed in Appendix D.

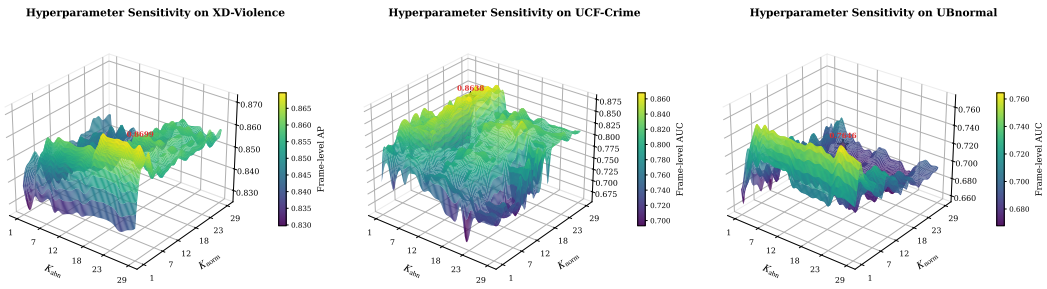
Table 13: Hyperparameter configurations for all three benchmarks. "—" indicates that the parameter is not applicable (SGP is disabled for UBnormal).

Dataset	K_N	K_A	α_G	β_{base}
XD-Violence	10	12	0.80	0.15
UCF-Crime	18	12	0.75	0.50
UBnormal	12	20	0.35	—

G.2 Prototype Count $(K_N,\ K_A)$

We evaluate the effect of varying prototype granularities. We first perform a coarse grid evaluation over $K_N, K_A \in \{1, 7, 12, 18, 23, 29\}$ on each dataset to identify high-performing plateaus, followed

by local refinement around candidate peaks (yielding the final configurations in Table 13, such as K_N=10 on XD-Violence and K_A=20 on UBnormal). Figure 11 shows the resulting coarse-grid heatmaps for XD-Violence (frame-level AP), UCF-Crime (frame-level AUC), and UBnormal (frame-level AUC).



(a) XD-Violence (AP) (b) UCF-Crime (AUC) (c) UBnormal (AUC)

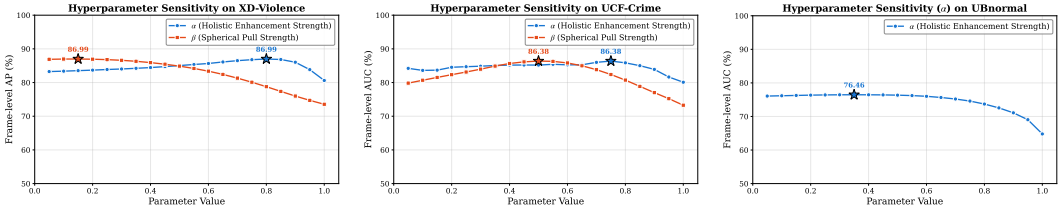
Figure 11: **Prototype count sensitivity.** Each cell shows the primary metric achieved by SphereVAD across varying counts of K_N normal prototypes (x-axis) and K_A anomalous prototypes (y-axis). Performance remains broadly stable across a wide interior region: on XD-Violence the top-quartile region (AP $\geq$ 0.860) covers roughly half the grid, and on UCF-Crime the AUC remains above 0.840 for all $K_N, K_A \geq 7$. Single-prototype baselines ($K_N = 1$ or $K_A = 1$) incur significant drops ($\sim$3–5%), confirming the utility of multi-prototype clustering on the hypersphere.

Observations.

- **Plateau behaviour.** For all three datasets, varying prototype counts within the central region $K_N, K_A \in [12, 23]$ yields relatively smooth performance with modest variance. Deviations in this interior region remain within manageable limits, showing that SphereVAD is tolerant to the specific choice of prototype partition size.
- **Single-prototype degeneration.** Restricting representations to a single prototype ($K_N = 1$ or $K_A = 1$) causes severe performance degradation, particularly on XD-Violence where visual manifestations of anomalous actions vary substantially across subclasses.

G.3 HSA Mixing Coefficient α_G and SLERP Base Pull Strength β_{base}

Figure 12 illustrates performance as a function of α_G and β_{base} when swept across $[0, 1]$ while holding remaining parameters fixed to their default settings. Because SGP is disabled for UBnormal, only the α_G curve is reported for that benchmark.



(a) XD-Violence (AP) (b) UCF-Crime (AUC) (c) UBnormal (AUC, α_G only)

Figure 12: **Sensitivity to α_G (Holistic Enhancement Strength) and β_{base} (Spherical Pull Strength).** Curves are obtained by sweeping one parameter over $[0, 1]$ while holding other variables fixed. Dashed horizontal lines indicate peak observed metrics (XD-Violence: AP = 86.99%; UCF-Crime: AUC = 86.38%; UBnormal: AUC = 76.46%). β_{base} is not applicable to UBnormal.

HSA mixing coefficient α_G. $\alpha_G \in [0, 1]$ governs the contribution of cross-video context relative to individual segment representations in Eq. (5).

- $\alpha_G = 0$ **(HSA disabled).** SphereVAD reverts to an M1+SGP pipeline without cross-video scene aggregation. The visible decline verifies the benefit of reinforcing segment consistency via global scene priors across untrimmed videos.
- **Stable plateau.** Under 1D sweeps, all three benchmarks exhibit stable trajectories within $\alpha_G \in [0.3, 0.8]$, with isolated metric changes remaining moderate.
- $\alpha_G \to 1$ **(pure attention).** Replacing intrinsic clip features completely with cross-video contextual aggregates degrades performance ($\sim$1–2%), as segment-specific temporal cues become over-smoothed.

SLERP base pull strength β_{base}. $\beta_{\text{base}} \in [0, 1]$ modulates the maximum displacement step during score-adaptive spherical interpolation (Eq. (44)).

- $\beta_{\text{base}} = 0$ **(SGP disabled).** Performance reverts directly to the M2 baseline, confirming that geodesic alignment plays an active role in resolving borderline segment scores.
- **Stable plateau.** Both XD-Violence and UCF-Crime display steady performance across $\beta_{\text{base}} \in [0.1, 0.5]$ in isolated 1D sweeps.
- $\beta_{\text{base}} \to 1$ **(excessive pull).** Excessive pull strength forces uncertain clips aggressively toward prototypes regardless of directional confidence. While the uncertainty factor $(1 - \frac{1}{2}\hat{d}_i)$ moderates the movement, extreme values ($\beta_{\text{base}} > 0.8$) lead to boundary oversaturation and performance loss.

G.4 Unified Hyperparameter Robustness

To evaluate the out-of-the-box generalization of SphereVAD, we test a *single, unified* default configuration across all three benchmarks: $K_N = 12$, $K_A = 18$, $\alpha_G = 0.5$, and $\beta_{\text{base}} = 0.5$ (with SGP disabled for UBnormal as established). These settings represent canonical median values observed across the parameter ranges. Table 14 contrasts this unified baseline against the per-dataset settings reported in Table 1.

Table 14: **Unified hyperparameter evaluation.** A single configuration ($K_N = 12$, $K_A = 18$, $\alpha_G = 0.5$, $\beta_{\text{base}} = 0.5$) evaluated across all three benchmarks. Δ denotes the difference in percentage points (pp). "—" indicates SGP is disabled.

Setting	**XD-Violence** AP(%)	**XD-Violence** AUC(%)	**UCF-Crime** AUC(%)	**UBnormal** AUC(%)
Per-dataset configuration	86.99	95.74	86.38	76.46
Unified default	85.65	95.10	83.42	72.31
Δ (pp)	-1.34	-0.64	-2.96	-4.15

Key observations include:

- **Robust baseline performance across benchmarks.** Under the single unified configuration, SphereVAD achieves 85.65% AP on XD-Violence, 83.42% AUC on UCF-Crime, and 72.31% AUC on UBnormal. On XD-Violence, this uncalibrated setting substantially surpasses prior training-free zero-shot methods (e.g., VADTree at 68.85% AP and PANDA at 70.16% AP). On UCF-Crime and UBnormal, although joint parameter shifts lead to moderate drops (-2.96 % and -4.15 pp, respectively), the unified model remains competitive, outperforming earlier training-free baselines such as MCANet (62.94% on UBnormal) and VADTree (65.80% on UBnormal).
- **Compound parameter shift.** While isolated sweeps along single dimensions exhibit broad plateau stability, simultaneously shifting multiple hyperparameters away from their empirical optima causes compounding geometric misalignments, which explains the observed performance gap.
- **Practical out-of-the-box utility.** This evaluation confirms practical applicability: even without dataset-specific tuning, SphereVAD provides strong zero-shot anomaly detection performance out of the box.

Summary. SphereVAD demonstrates consistent behavior across reasonable variations in prototype granularity, attention weighting, and geodesic displacement. The unified baseline delivers competitive zero-shot performance across distinct surveillance domains, showing that the core benefits stem from the underlying spherical geometric formulations rather than narrow parameter assignments.

G.5 Downstream Extension: Adaptability to Weakly-Supervised MIL

To investigate whether the geometric structures unlocked by spherical centering can benefit traditional learning paradigms, we evaluate SphereVAD representations under the standard weakly-supervised video anomaly detection (WS-VAD) protocol on **ShanghaiTech** [30, 55] (238 training and 199 testing videos). Comparative results against state-of-the-art methods are reported in Table 4 of the main text.

Feature Centering. Following the unified formulation in Sec. 3.2, we pool intermediate representations (layer 31 of the frozen Qwen3.5 backbone) from both the synthetic calibration set and the ShanghaiTech training split to compute the unified Fréchet mean $\boldsymbol{\mu}_{\text{unified}}$ via Karcher iteration (Eq. (1)). All features $\hat{\mathbf{f}}^l$ are projected onto $\mathcal{S}^{D-1}$ via Eq. (2) using this reference point.

Lightweight MIL Head Architecture and Training. We treat each untrimmed video as a bag of T temporal clips. The downstream classifier consists of a 2-layer temporal MLP: a Linear layer mapping from $D = 4096$ to 512 followed by ReLU and Dropout ($p = 0.5$), and a final Linear projection from 512 to 1 with a Sigmoid activation yielding clip-level anomaly probabilities $s_i \in [0, 1]$. Following the robust MIL formulation [43], during training we select the top-k ($k = \lfloor T/16 \rfloor + 1$) highest-scoring clips per video to compute the bag-level prediction. The head is optimized via the standard binary cross-entropy bag loss paired with temporal smoothness and sparsity constraints [40]. We train the head using AdamW with an initial learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and batch size of 32 for 50 epochs on a single GPU.

Ablation on Spherical Centering. To verify that the downstream performance gains originate from the unfolded manifold structure rather than the classifier's capacity, we train the identical MIL head on raw features that are merely ℓ_2-normalized without spherical centering. As shown in Table 15, removing spherical centering causes a notable drop of **3.53%** in frame-level AUC. This confirms that eliminating conical collapse fundamentally facilitates downstream boundary learning.

Table 15: Ablation of spherical centering on downstream MIL learning (ShanghaiTech, frame-level AUC %).

Feature Representation	**ShanghaiTech AUC (%)**
Frozen Qwen3.5 (ℓ_2-normalized only) + MIL	92.85
Frozen Qwen3.5 (**Spherically Centered**) + MIL	**96.38** (**+3.53**)

H Reproducibility Statement & Symbol Table

H.1 Full Symbol Table

Table 16: Summary of mathematical symbols used throughout the paper and appendix.

Symbol	Description
f	Raw intermediate-layer feature
$\tilde{f}$	ℓ_2-normalised feature on $\mathcal{S}^{D-1}$
$\hat{f}$	Spherically centered feature
f^l	Main feature (last token hidden state)
f^v	Visual feature (visual-last token hidden state)
D	Feature dimensionality
$\mathcal{S}^{D-1}$	Unit hypersphere in $\mathbb{R}^D$
$\boldsymbol{\mu}_{\text{unified}}$	Unified Fréchet mean
$\boldsymbol{\mu}_k^{\text{norm}}$, $\boldsymbol{\mu}_k^{\text{abn}}$	Normal / anomalous vMF prototypes
K_N, K_A	Number of normal / anomalous prototypes
κ	vMF concentration parameter
α_G	HSA mixing coefficient
β_{base}	SLERP base pull strength
β_i	Score-adaptive pull strength for clip i
s	Anomaly score $\in [0, 1]$
$d_{\text{geo}}(\cdot, \cdot)$	Geodesic distance on $\mathcal{S}^{D-1}$
$\text{Log}_{\boldsymbol{\mu}}(\cdot)$	Riemannian logarithmic map at $\boldsymbol{\mu}$
$\text{Exp}_{\boldsymbol{\mu}}(\cdot)$	Riemannian exponential map at $\boldsymbol{\mu}$
$\text{Slerp}(\cdot, \cdot, \cdot)$	Spherical linear interpolation
$\mathbf{A}$	Cross-video sparse attention matrix (HSA)
$\mathbf{A}_H$	Intra-video sparse attention matrix (SGP)
ρ_{low}, ρ_{high}	MAD-based ambiguity interval bounds
$\boldsymbol{\mu}_{\text{dom}}^c$	Dominant prototype for class c
$\mathcal{D}_{\text{syn}}$	Synthetic calibration dataset
M	Number of synthetic calibration images
N	Number of test videos
T_i	Number of clips in video V_i

H.2 Random Seed & Determinism

SphereVAD is designed to be deterministic given a fixed random seed. A global seed (default: `42`) controls the only stochastic component in the pipeline—spherical K-Means initialisation—via `sklearn.cluster.KMeans(random_state=42)` with $n_{\text{init}} = 10$. All remaining operations (Fréchet mean, spherical centering, HSA, vMF scoring, SGP, Gaussian smoothing) are closed-form deterministic computations. We additionally set `torch.backends.cudnn.deterministic=True` and `torch.backends.cudnn.benchmark=False` to ensure reproducible MLLM forward passes. We verified stability by repeating the full pipeline with multiple seeds $\{42, 123, 456, 789, 1024\}$ on the same hardware and observed negligible variance across runs.

H.3 Code & Data Availability

The complete source code—including feature extraction, the geometric inference pipeline, evaluation scripts, and configuration files—is publicly available at `https://github.com/S1inetzz/SphereVAD`. Pre-extracted features for all three benchmarks and the synthetic calibration set will also be provided, enabling full reproduction of the geometric inference pipeline *without* GPU access.

The synthetic calibration data generation scripts and meta-prompts are included in the repository for reproducibility. We emphasise that the synthetic data are *not* a contribution of this work; they serve solely as directional calibrators for vMF prototype construction and can be replaced by any alternative set of normal/abnormal reference images. All synthetic images are machine-generated and contain no real individuals or personally identifiable information.

All four MLLM backbones (Qwen3.5, Qwen3-VL, InternVL3, LLaVA-OV-1.5) are publicly available on Hugging Face and used in their frozen, unmodified states. All three evaluation benchmarks (XD-Violence, UCF-Crime, UBnormal) are publicly available and used with their standard test splits under their respective licences.

I Limitations

While SphereVAD achieves strong training-free performance, several limitations merit discussion. **Dependence on MLLM feature quality.** SphereVAD does not learn new representations; it operates entirely on frozen intermediate-layer features. Consequently, its performance ceiling is bounded by the anomaly-discriminative capacity already latent in the backbone. As shown in Table 3, weaker backbones (e.g., LLaVA-OV-1.5) yield noticeably lower results, indicating that the framework amplifies but cannot create discriminability absent from the features. **Short-video limitation of SGP.** The intra-video Spherical Geodesic Pulling mechanism relies on MAD-based tri-classification and neighbour consensus, both of which require sufficient temporal context. For very short videos (e.g., UBnormal, averaging $\sim$12 seconds per video), SGP provides no benefit and is disabled (Table 2, M2 = M3 on UBnormal). Extending SGP to short-clip scenarios—e.g., via cross-video pulling or temporal padding—remains future work. **Synthetic calibration domain gap.** Although unified Fréchet mean centering effectively absorbs the $\sim 5^\circ$ rotational bias between synthetic and real domains, a residual domain gap persists. The current synthetic generation pipeline (Appendix B) uses generic meta-categories; domain-adaptive calibration strategies that tailor synthetic data to specific deployment environments could further improve performance. **Single-anomaly-type assumption in SGP.** The dominant prototype voting mechanism (Appendix D.2) assumes that each video contains at most one prevailing anomaly type. Videos with multiple co-occurring anomaly categories (e.g., simultaneous arson and assault) may not be optimally handled, as the single dominant anomalous prototype cannot represent both event types. **Quadratic cost of HSA.** The cross-video Holistic Scene Attention module constructs a pairwise similarity matrix over all test clips, incurring $O(N_{\text{total}}^2 D)$ cost. While manageable for current benchmarks ($\sim$97K clips, $<$10 s on CPU), scaling to deployment scenarios with millions of clips would require approximate nearest-neighbour indexing or hierarchical attention schemes.

J Broader Impact

Positive societal impact. SphereVAD is designed to enhance public safety by enabling rapid, training-free deployment of video anomaly detection in surveillance systems. Its zero-shot nature eliminates the need for target-domain data collection and annotation, which is particularly valuable for under-resourced communities and emergency response scenarios where labelled data are unavailable. The training-free design also substantially reduces the computational carbon footprint compared with methods requiring GPU-intensive training. **Potential risks.** As with all surveillance-related technologies, SphereVAD carries inherent risks of misuse, including unwarranted mass surveillance, privacy violations, and disproportionate targeting of marginalised populations. The system may inherit biases present in the pre-trained MLLM backbone (e.g., demographic biases in visual recognition) or in the synthetic calibration data (e.g., over-representation of certain anomaly scenarios). False positives could lead to unwarranted interventions, while false negatives could result in missed genuine threats. **Mitigation strategies.** We advocate for the following safeguards: (i) deployment should comply with applicable privacy regulations and be subject to institutional oversight; (ii) anomaly scores should be treated as alerts requiring human verification rather than automated decisions; (iii) regular auditing of detection outcomes across demographic groups should be conducted to identify and mitigate bias; (iv) the synthetic calibration data should be reviewed to ensure diverse and balanced representation of scenarios. **Dual-use considerations.** The geometric inference framework itself (spherical centering, vMF scoring, geodesic pulling) is domain-agnostic and could be applied to beneficial tasks beyond surveillance, such as medical anomaly detection, industrial quality control, or environmental monitoring. We release the methodology with the intent of advancing scientific understanding of training-free inference on hyperspherical manifolds.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state four contributions (training-free vMF geodesic inference framework, unified Fréchet mean centering with theoretical guarantee, HSA and SGP modules, and SOTA training-free results on three benchmarks). All claims are supported by theoretical analysis (Propositions 1–2 in Section 3.2 and Appendix A) and experimental results (Tables 1–3, Figures 3–4 in Sections 4.1–4.5). The scope is appropriately limited to training-free VAD and the paper does not overclaim competitiveness with the best fully supervised methods.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section (Appendix I) discusses five specific limitations: dependence on MLLM feature quality, short-video limitation of SGP, synthetic calibration domain gap, single-anomaly-type assumption in SGP, and quadratic cost of HSA. Each limitation is accompanied by concrete evidence (e.g., Table 3 for backbone dependence, Table 2 for UBnormal SGP limitation) and directions for future work.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Propositions 1 and 2 are formally stated in Section 3.2 with explicit assumptions (vMF class distributions, mixture weights). Complete proofs are provided in Appendix A.2 and A.3, including a supporting Lemma 1 (Appendix A.2) with a separate proof. The Karcher iteration convergence guarantee (Theorem 1) is stated with full conditions in Appendix A.1.3. All results are numbered and cross-referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the complete prompt template (Appendix C.1), DLSP layer selection procedure and selected layers for all backbones (Appendix C.3), full hyperparameter settings (Table 13), the complete SGP algorithm (Algorithm 1), the synthetic data generation pipeline (Appendix B), and hardware specifications (4×A100 GPUs). All four MLLM backbones are publicly available on Hugging Face. Random seeds and determinism settings are specified in Appendix H.2.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code and reproduction instructions are publicly available at `https://github.com/S1inetzz/SphereVAD`. Code and data availability are described in Appendix H.3.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4 (Implementation Details) specifies the backbone (frozen Qwen3.5), extraction layer (layer 31 via DLSP), calibration set size (∼2,000 images), and hardware (4×A100). Appendix G provides the complete hyperparameter table (Table 13), sensitivity analyses (Figures 11 and 12), and a unified-setting evaluation (Table 14). The training details for the weakly supervised extension are provided in Appendix G.5.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars or confidence intervals for the main results. However, as noted in Appendix H.2, the pipeline is designed to be deterministic (fixed random seed for K-Means, deterministic CUDA settings), and the authors verified negligible variance across five different seeds $\{42, 123, 456, 789, 1024\}$. The only stochastic component is spherical K-Means initialisation with `n_init=10`, which produces near-identical results across runs.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E provides feature extraction times for both 1×A100 and 4×A100 configurations (Table 9), per-stage geometric inference times on a single CPU core (Table 10), storage costs, online throughput (59.3 FPS on 1×A100), and comparison with existing methods (Table 11). The GPU type (A100-80GB) and CPU type (Intel Xeon Platinum 8375C) are specified.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The synthetic calibration data contain no real individuals or personally identifiable information (Appendix H.3). No depictions of minors are permitted in the synthetic generation pipeline (Appendix B.1). All evaluation benchmarks are publicly available and used under their respective licences. The paper discusses potential risks and mitigation strategies in the Broader Impact section (Appendix J).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix J provides a dedicated Broader Impact discussion covering: positive impacts (enhanced public safety, reduced computational footprint, accessibility for under-resourced communities), potential risks (mass surveillance, privacy violations, demographic bias, false positives/negatives), specific mitigation strategies (human verification, demographic auditing, privacy regulation compliance), and dual-use considerations.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: SphereVAD does not release a new pre-trained model, image generator, or scraped dataset. The framework operates on frozen, publicly available MLLM backbones and synthesises calibration images containing no real individuals. The released code will implement a geometric inference pipeline (no trainable parameters), which poses negligible standalone misuse risk beyond the existing MLLMs it depends on.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All four MLLM backbones, Qwen3.5 [41], Qwen3-VL [4], InternVL3 [57], and LLaVA-OneVision-1.5 [3], are cited with their original publications and are publicly available on Hugging Face. The evaluation benchmarks XD-Violence [45], UCF-Crime [40], and UBnormal [2] are also cited. Appendix H.3 discusses asset availability and usage.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The synthetic calibration dataset ($\sim$2,000 images) is documented in Appendix B, including the generation pipeline (Appendix B.1), data format and pairing schema (Appendix B.2, Table 5), sample images (Figure 6), sample size justification (Appendix B.4), and sub-category taxonomy and comparison (Tables 6 and 7).

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects. All data are either from existing public benchmarks or synthetically generated via LLM and text-to-image models.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects research. No IRB approval is required.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper explicitly describes two uses of LLMs that are integral to the methodology: (1) frozen MLLMs (Qwen3.5, etc.) serve as the feature extraction backbone whose intermediate-layer representations are the input to the geometric inference pipeline (Section 3.2); (2) Claude Sonnet 4.5 is used to generate structured scene descriptions for the synthetic calibration dataset (Appendix B.1, Stage 1), and Gemini 2.5 Flash is used for image synthesis (Stage 2). Both uses are thoroughly documented with specific model names, access methods (API), and their roles in the pipeline.