BRIDGING SHANNON SURPRISAL AND LATENT GEOMETRY: A UNIFIED FRAMEWORK FOR VLM STORAGE, ATTENTION ROUTING, AND DATA PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As Vision-Language Models (VLMs) expand into unconstrained, in-the-wild continuous spatiotemporal domains, uniform vector processing creates severe KV-Cache and computational attention bottlenecks. We propose a sparsity-driven methodology that bridges these geometric latent constraints with Shannon's Information Theory. By modeling dense background representations as a von Mises-Fisher (vMF) distribution, we derive that local surprisal exhibits an affine relationship with a vector's cosine similarity with the inverse global mean ($-\bar{v}$). To overcome representation degeneration—where vectors are constrained within narrow anisotropic boundaries—we formulate an exponential weighting function equipped with a Dimensionality Compensator (α) and a Structural Domain Shifter (β). This formulation drives a dual-phase optimization framework. First, in static environments, our Surprisal-Weighted Mixed-Precision (SWMP) strategy achieves a 57.50% memory reduction by quantizing low-surprisal representations. Second, for continuous video streams, we introduce Temporal SWMP, leveraging an Exponential Moving Average (EMA) to transition the framework into an Information-Theoretic Attention Router. We further observe strong metric coupling between adjacent-difference routing and a second-difference kinematic evaluation proxy. Under extreme Iso-KV Cache constraints ($\sim$3% of frames), Temporal SWMP exceeds uniform sampling in recall while exhibiting selection behavior distinct from adjacent-difference routing. Together, these results support an information-theoretic criterion for attention routing in resource-constrained continuous VLM deployments across unpredictable real-world environments.

1 INTRODUCTION

The deployment of large-scale Vision-Language Models (VLMs), such as CLIP, in unconstrained spatiotemporal environments (e.g., autonomous driving, continuous video) has exposed a severe computational bottleneck Radford et al. (2021); Jia et al. (2021); Alayrac et al. (2022); Wang et al. (2024). Continuous data streams are overwhelmingly dominated by redundant background frames. Allocating uniform KV-Cache storage and attention bandwidth to these ubiquitous features induces heavy "attention dilution," severely degrading system scalability and real-time responsiveness.

While prior studies have independently explored information-theoretic data pruning Tishby et al. (1999); Kullback & Leibler (1951); Cover & Thomas (2006); MacKay (2003), hypersphere directional statistics Mardia & Jupp (1999); Davidson et al. (2018), and the topological flaw of representation degeneration Gao et al. (2019); Ethayarajh (2019); Wang & Isola (2020), they have remained structurally siloed. This paper bridges this critical gap by introducing a sparsity-driven weighting methodology that uniquely translates Shannon's Information Theory Shannon (1948) into the exact geometry of high-dimensional latent spaces. We mathematically demonstrate that the Shannon Surprisal of any local vector exhibits an affine relationship with its cosine similarity with the inverse global mean ($-\bar{v}$).

However, translating this theoretical surprisal into an exponential weighting function encounters a critical topological roadblock: representation degeneration (anisotropy) in high-dimensional spaces ($d \geq 512$). Because local vectors cluster tightly within a narrow degenerate cone, raw cosine sim-

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

ilarities can occupy a bounded negative range, causing aggressive exponential scaling to approach finite-precision underflow. To fundamentally bypass this, our formulation integrates a dimensionality compensator (α) to dynamically stretch infinitesimal angular variances, and a domain shifter (β) to offset the strictly negative geometric bounds. This architectural safety net guarantees that global latent sparsity is exponentially rewarded without numerical collapse.

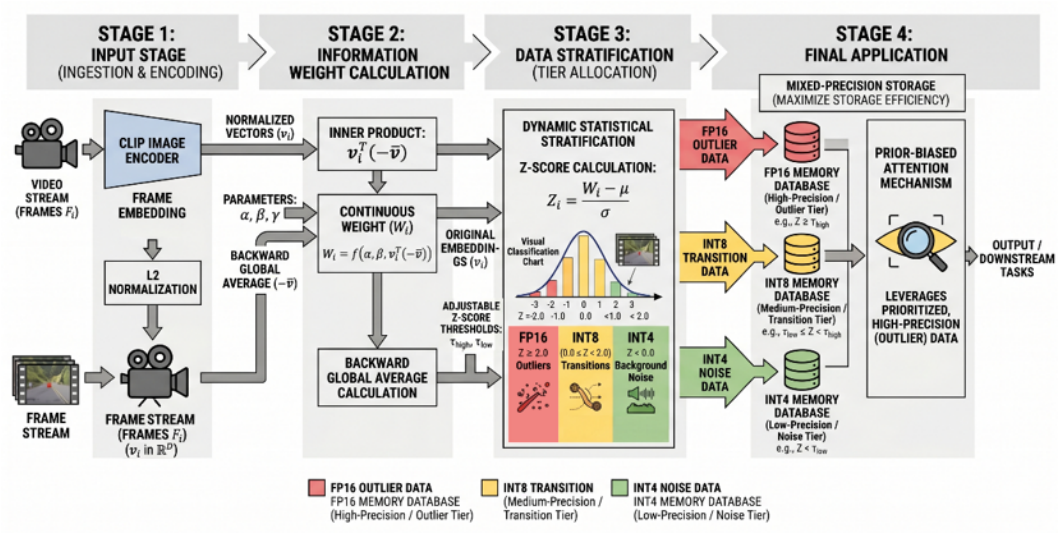


Figure 1: The overall conceptual pipeline of the proposed Latent-Geometric Surprisal-based visual data processing workflow. (Stage 1) A continuous video frame stream is ingested and encoded into high-dimensional normalized vectors via a CLIP image encoder. (Stage 2) The system calculates the continuous information weight by computing the geometric inner product with the backward global average and applying structural parameters. (Stage 3) Through dynamic statistical stratification (Z-score calculation), the weights are allocated into three distinct precision tiers. (Stage 4) Finally, the stratified data is distributed into mixed-precision storage to maximize memory efficiency, while high-precision outlier data structurally dominates computational focus via the Prior-Biased Attention Mechanism.

Based on our mathematically bounded surprisal formulation, we propose a unified framework with three primary contributions:

- **Surprisal-Weighted Mixed-Precision (SWMP):** A strategic storage paradigm that quantizes mathematically redundant background noise while preserving high-surprisal anomalies in full precision, yielding a 57.50% memory footprint reduction, yet reveals a critical trade-off: extreme quantization (e.g., INT4) of background priors distorts the latent Riemannian manifold, inducing severe downstream semantic degradation.
- **Prior-Biased Attention Mechanism:** Explicitly injects continuous surprisal weights into the Transformer's self-attention layers as an inductive prior. Backed by rigorous mathematical proofs, this mechanism actively overrides representation degeneration.
- **Semantic Singularity Criterion & Temporal Routing:** Extending to continuous video streams under extreme Iso-KV Cache constraints ($\sim$3% of frames), Temporal SWMP functions as an Information-Theoretic Attention Router. It reveals strong metric coupling between adjacent-difference routing and kinematic evaluation proxies, while exhibiting a distinct selection pattern under constrained budgets.

2 Preliminaries and Related Work

2.1 Latent Geometry and Directional Statistics

Modern Vision-Language Models (VLMs) like CLIP Radford et al. (2021) map multimodal inputs into a shared L_2-normalized latent space, forming a d-dimensional Riemannian hypersphere

($\mathcal{S}^{d-1}$). However, these high-dimensional spaces fundamentally suffer from representation degeneration Gao et al. (2019); Ethayarajh (2019). Local vectors cluster tightly within a narrow anisotropic cone, where the global mean ($\bar{v}$) encapsulates generic, redundant background features. To robustly model probability distributions on $\mathcal{S}^{d-1}$, directional statistics frequently employ the von Mises-Fisher (vMF) distribution Mardia & Jupp (1999). In this geometric context, distant outliers residing toward the exact opposite direction ($-\bar{v}$) naturally capture rare, discriminative information.

2.2 KV-Cache Compression and Token Pruning

To alleviate the extreme computational and memory bottlenecks of attention mechanisms in continuous streams, conventional systems rely on token pruning or KV-Cache eviction strategies. Techniques such as Token Merging (ToMe) Bolya et al. (2023), Heavy-Hitter Oracle (H2O) Zhang et al. (2023), and StreamingLLM Xiao et al. (2024) attempt to dynamically drop or compress tokens based on heuristic attention scores, adjacent pixel similarities, or activation magnitudes. While hardware-level mixed-precision Dettmers et al. (2022) further reduces memory footprints, these kinematic and magnitude-based heuristics critically overlook the inherent information-theoretic value structurally encoded within the global geometric distribution of the latent space.

2.3 Shannon's Information Theory and Latent-Geometric Surprisal

In Shannon's Information Theory Shannon (1948), the information content of an event is inversely proportional to its probability, defined mathematically as Surprisal:

$$I(x) = -\log P(x)$$

Existing sparsity methods treat $P(x)$ empirically based on frequency or magnitude. In contrast, bridging this probabilistic surprisal directly into the continuous geometric space (using the alignment with $\bar{v}$) remains an unexplored frontier. Establishing a rigorous mathematical link between $I(x)$ and geometric topology is precisely the theoretical foundation required to transform heuristic routing into deterministic structural attention prioritization.

3 Information-Theoretic Methodology

This section establishes the mathematical foundations of our sparsity-driven weighting methodology. We constructively formulate the topological behavior of continuous spatiotemporal data on high-dimensional hyperspheres. By applying the Maximum Entropy Principle Jaynes (1957), we derive an exact geometric formulation for Shannon Surprisal, which we subsequently translate into a robust adaptive storage pipeline.

3.1 Topological Foundations and the MaxEnt Principle

Modern VLMs map multimodal inputs into a shared latent space, formally defined as a d-dimensional Riemannian hypersphere ($\mathcal{S}^{d-1}$) (Def. 1). Because discrete physical frames asymptotically converge to a continuous latent trajectory $v(t)$ (Def. 2), highly localized topological clusters naturally emerge in continuous streams, representing redundant semantic backgrounds (Axiom 1). To robustly define the geometric center of this continuous background, we utilize the Fréchet mean ($\bar{v}$) (Propisition 1), which uniquely minimizes geodesic distances along the curved surface (Remark 1). As demonstrated in Appendix A.2, for online operation, we approximate the evolving background direction using an L_2-Normalized EMA (N-EMA), requiring $O(d)$ computation per frame.

Having established $\bar{v}$ as the background prior, we apply Jaynes' Maximum Entropy Principle Jaynes (1957). As rigorously proven in Appendix A.3 (Theorem 1), the unique probability distribution $p(v)$ that maximizes Shannon entropy for this hyperspherical stream is the von Mises-Fisher (vMF) distribution. This leads directly to our foundational geometric substitution:

Lemma 1 (Affine Relationship of Shannon Surprisal). Given a vMF background model with mean direction $\bar{v}$ and concentration $\kappa > 0$, the informational surprisal of a normalized latent vector v_i is

$$I(v_i) = \kappa v_i^\top (-\bar{v}) - \log C_d(\kappa).$$

Thus, Shannon surprisal is an affine function of cosine similarity to the inverse background direction. For a fixed background direction and concentration,

$$I(v_i) - I(v_j) = \kappa \left[v_i^\top(-\bar{v}) - v_j^\top(-\bar{v}) \right],$$

Consequently, surprisal and inverse-background cosine similarity induce the same ordering over observations. *(The full formal definitions, axioms, and the Fréchet mean proposition are detailed in Appendix A.1).*

3.2 Exponential Influence Formulation and Structural Adaptation to Anisotropy

To translate the theoretical latent-geometric surprisal into a tangible metric for hardware memory allocation and downstream model attention, we define a continuous, sparsity-driven exponential weighting function $W(t)$. For practical hardware computation, this field is sampled at discrete frame intervals to yield W_i:

$$W_i = \exp(\alpha(v_i^\top(-\bar{v}) + \beta) + \gamma_b)$$

High-dimensional latent spaces ($d \geq 512$) frequently suffer from representation degeneration (anisotropy). This topological flaw confines local vectors within a narrow degenerate cone, predominantly bounding raw cosine similarities within the negative domain. To ensure the exponential weighting formulation functions effectively within these geometric constraints without triggering numerical underflow, we operationalize three structural parameters:

- α **(Dimensionality Compensator):** Acts as a temperature scalar that stretches infinitesimal, microscopic angular variances between sequential frames in high-dimensional spaces, amplifying the distinction between common noise and sparse anomalies.
- β **(Domain Shifter):** A critical safety parameter that offsets the inherently negative geometric bounds caused by representation degeneration. By translating the bounded score range within the exponent, β calibrates the numerical operating range of the exponential weights and mitigates finite-precision underflow or overflow under aggressive scaling.
- γ_b **(Baseline Shift):** Acts as a fixed additive offset that controls the baseline numerical scale of the exponential weights.

Governed by the anisotropic constraints of the latent space—where the raw inner product is strictly bounded by $v_i^\top(-\bar{v}) \leq \delta_{max} < 0$—this weighting function stratifies representations into three distinct stages of computational influence. In the **(1) Common Baseline Stage**, vectors clustered around the mean yield minimal surprisal, driving W_i toward its lowest bound. During the **(2) Transition Phase**, as vectors approach the geometric limit, β shifts the exponent toward a numerically resolvable operating range. Finally, in the **(3) High Sparsity Phase**, positively shifted outliers residing at the extreme boundaries of the degenerate cone are exponentially amplified by α, structurally dominating computational focus.

Practically, these discrete surprisal weights W_i directly drive a Surprisal-Weighted Mixed-Precision (SWMP) storage paradigm. By stratifying vectors into discrete precision tiers based on their computed influence—retaining sparse anomalies in full precision (e.g., FP16) to preserve critical discriminative semantics, while quantizing redundant backgrounds (e.g., INT4)—SWMP is used for massive memory reductions while protecting the integrity of rare events.

4 Theoretical Analysis of Prior-Biased Attention

Having geometrically derived the continuous latent-geometric surprisal $W(t)$ and its discrete hardware sampling W_i, this section formalizes how this continuous information metric is integrated into the downstream Transformer architecture. We theoretically prove that injecting this metric mathematically acts as a Bayesian Prior-Biased Attention mechanism, driving Asymptotic Attention Monopoly toward critical anomalies while structurally avoiding numerical instability.

4.1 Prior-Biased Logit Intervention

The standard self-attention mechanism Vaswani et al. (2017); Beltagy et al. (2020) computes the contextual mapping of a sequence by calculating the scaled dot-product between a query q and

historical keys k_j. The attention weight a_j assigned to the j-th frame is obtained via the softmax function:

$$a_j = \text{softmax}(s_j) = \frac{\exp(s_j)}{\sum_{m=1}^{N} \exp(s_m)}$$

where $s_j = \frac{q^\top k_j}{\sqrt{d}}$ represents the compatibility logit. Relying solely on this compatibility is limiting in continuous spatiotemporal environments, where visual similarity distributes computational bandwidth across redundant environmental noise.

To prioritize sparse vectors through an explicit logit-level intervention, we introduce our information-theoretic surprisal directly into the Transformer's self-attention mechanism. A naive approach of multiplying the inputs directly is susceptible to magnitude neutralization during standard Layer Normalization. Therefore, we introduce it as an inductive prior inside the pre-softmax logit.

Lemma 2 (Prior-Biased Logit Injection): Injecting an exponential geometric prior into the attention mechanism translates to an additive shift within the logit space. By defining a prior factor based on our continuous surprisal weight $g_{\text{prior}}(k_j) \propto \exp(\lambda W_j)$, where λ is a confidence scalar, the modified Attention equation becomes:

$$a_j = \frac{\exp(s_j + \lambda W_j)}{\sum_{m=1}^{N} \exp(s_m + \lambda W_m)}$$

This logit-level integration provides a formalized method for attention redistribution. The modified pre-softmax attention score is formalized as $e_{i,j} = \frac{q_i k_j^\top}{\sqrt{d}} + \lambda W_j$, and consequently, the global attention computation is redefined as:

$$\text{Attention}(Q, K, V, W) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \lambda W_{\text{bias}}\right) V$$

This demonstrates that an additive bias ($+\lambda W_j$) directly inside the softmax logit serves as a principled solution to inject informational prioritization into the Transformer architecture.

4.2 Asymptotic Attention Monopoly

Having established the prior-biased logit formulation, we now analyze the asymptotic behavior of the modified attention distribution under spatiotemporal redundancy.

Theorem 2 (Asymptotic Attention Monopoly): Suppose a temporal context window consists of one sparse anomaly trajectory v_{rare} and multiple common background frames v_c. If the anomaly has a unique maximum prior weight,

$$W_{\text{rare}} - \max(W_c) > 0,$$

then as the prior-strength scalar $\lambda \to \infty$, the attention mass allocated to the anomaly converges to one ($a_{\text{rare}} \to 1$).

Theorem 2 shows that sufficiently increasing the prior strength asymptotically dominates fixed differences in the compatibility logits and concentrates attention on the unique highest-weight representation. We refer to this limiting concentration behavior as **Attention Monopoly**.

4.2.1 Geometric Bounds and the Finite-Precision Role of the Domain Shifter (β)

While Theorem 2 characterizes Attention Monopoly in exact arithmetic, practical implementations evaluate the exponential weight under finite numerical precision. Due to representation degeneration, models such as CLIP can confine inverse-background similarities to a bounded negative interval.

Definition 3 (Anisotropic Cone Constraint): In a degenerated multimodal latent space, the continuous vector field is geometrically confined within a narrow anisotropic cone with maximum angular spread $\theta_{\max} < \frac{\pi}{2}$. **Lemma 3 (Strict Negative Bound of Inverse Similarity):** Under the

Anisotropic Cone Constraint, the inverse-background similarity

$$s_i = v_i^\top(-\bar{v})$$

remains in the negative domain and is bounded above by $\delta_{\max} < 0$.

Theorem 3 (Finite-Precision Operating Range of the Domain Shifter): Let $s_i \in [\delta_{\min}, \delta_{\max}]$ denote the observed inverse-background score range, and let $[\tau_{\min}, \tau_{\max}]$ denote the effective log-domain range supported by the numerical implementation of the exponential weighting function

$$W_i = \exp(\alpha(s_i + \beta) + \gamma_b)\,.$$

For $\alpha > 0$, the full observed exponent range remains within the representable interval whenever

$$\tau_{\min} \le \alpha(\delta_{\min} + \beta) + \gamma_b \quad \& \quad \alpha(\delta_{\max} + \beta) + \gamma_b \le \tau_{\max}.$$

Thus, β shifts the numerical operating range of the exponential weights without changing the ordering of the underlying geometric scores.

Theorem 3 characterizes β as a finite-precision domain shifter rather than a prerequisite for the asymptotic concentration of Theorem 2. In practical low-precision implementations, this shift provides direct control over underflow and overflow when aggressive exponential scaling is applied.

(Note: While our methodology primarily targets inference-time KV-Cache bottlenecks, Appendix A.5 provides a local gradient analysis of the differentiable prior intervention.)

4.3 Mechanism of Attention Redistribution

The logit-level integration defined in our Prior-Biased Attention provides a method for attention redistribution. Because the softmax function exponentially amplifies logit differences, vectors with high surprisal weights structurally capture massively disproportionate portions of the probability distribution. As derived in our theoretical analysis, the prior-induced attention odds multiplier (N) between a rare anomalous vector and a common background vector is defined by the exponential scaling factor:

$$N = \exp(\lambda(W_{\text{rare}} - W_{\text{common}})) = \exp(\lambda \Delta W)$$

To illustrate the practical scaling of this mechanism in a continuous temporal environment, assume an empirically optimized baseline weight gap of $\Delta W = 0.10$:

- If $\lambda = 10$, the prior multiplies the anomaly-to-background attention odds by $N \approx 2.71$.
- If $\lambda = 50$, this odds multiplier increases to $N \approx 148.4$.
- If $\lambda = 100$, the prior multiplies the anomaly-to-background attention odds by $N \approx 22,026$, strongly favoring the high-surprisal representation.

In a continuous video stream comprising numerous redundant elements, the Prior-Biased Attention framework ensures that highly sparse vectors receive exponentially scaled priority. Rather than claiming behavioral dominance across all unverified downstream tasks, this structural property serves as a foundational mechanism to mitigate attention dilution. By flattening redundant noise, it establishes an optimal cognitive safety net, ensuring systems allocate computational resources effectively to preserve critical semantic shifts within noisy spatiotemporal environments.

5 Experimental Framework and Results

To empirically validate the proposed sparsity-driven weighting methodology and the prior-biased attention mechanism, we designed a comprehensive two-phase experimental framework. Phase I focuses on the static representation space, rigorously reverse-engineering the geometric properties of the CLIP latent space to cross-validate our theoretical bounds. Building upon these structural insights, Phase II extends the framework into the spatiotemporal domain, demonstrating our methodology as a dynamic Attention Router to mitigate KV-Cache bottlenecks in continuous video streams.

Phase I: Static Representation and Geometric Validation

5.1 Geometric Reverse-Engineering and Validation of Bounds

For the static phase, we extracted 53,926 high-dimensional foundational embeddings via a pre-trained CLIP model (ViT-L/14) on the MS-COCO dataset Lin et al. (2014). By evaluating the raw surprisal weights without the domain shifter ($\beta = 0$), we reverse-engineered the dataset's latent space geometry. An ablation study revealed that setting the dimensionality compensator to $\alpha = 1$ yielded a mathematically narrow weight variance, fundamentally failing to distinguish anomalies. In contrast, $\alpha = 30$ successfully amplified global latent sparsity. Under this configuration, the cosine similarities between all local vectors and the inverse mean ($-\bar{v}$) were strictly confined between -0.850 and -0.197. This provides stark empirical evidence of representation degeneration (anisotropy), perfectly cross-validating our theoretical bound of $\delta_{max} = -0.197$.

To evaluate the Prior-Biased Attention mechanism under these constraints, we measured the Attention Monopoly Ratio (AMR). Under a Uniform Baseline ($\lambda = 0$), the sparse anomaly captured a probability mass of roughly 0.0018%. By applying Prior-Biased Attention ($\lambda = 100, \beta = 0$), the AMR exhibited a strictly positive shift to 0.0024%. Crucially, this empirical scaling matches the theoretical attention multiplier prediction, $N \approx 1.31$, confirming the analytical relationship between the surprisal-induced weight gap and attention redistribution. The relatively small perturbation at $\beta = 0$ further motivates the domain shift as a finite-precision scale-control mechanism: β relocates the exponential operating range without altering the underlying ordering of geometric novelty.

5.2 Surprisal-Weighted Stratification Mechanics

Having validated the structural parameters, we applied the Surprisal-Weighted Mixed-Precision (SWMP) strategy. Vectors were stratified into discrete precision tiers based on their computed latent-geometric surprisal weights (W_i):

- **High-Precision (FP16, Top 5%):** Peak surprisal vectors retained to preserve discriminative semantics.
- **Intermediate (INT8, Mid 55%):** Transition-phase vectors undergoing moderate quantization.
- **Low-Precision (INT4, Bottom 40%):** Statistically predominant background noise.

The baseline uniform FP16 storage for the dataset required 78.99 MB. By implementing SWMP, the total memory footprint was drastically reduced to 33.57 MB, achieving a massive 57.50% reduction in storage capacity, though at the cost of manifold curvature collapse (detailed in Appendix C.2), which motivates the temporal routing in Phase II.

Phase II: Dynamic Attention Routing in Spatiotemporal Streams

To transition from a static sandbox to real-world spatiotemporal environments exhibiting extreme temporal redundancy, we evaluated our methodology on continuous video sweeps (16,095 frames at 12Hz) from the nuScenes dataset Caesar et al. (2019).

5.3 Dynamic Statistical Stratification and Kinematic-Proxy False Positives

Relying on a static threshold enforces a uniform sensitivity, which can be suboptimal for properly capturing vectors that deviate from the localized central meaning. To operationalize our methodology dynamically, Temporal SWMP substitutes the static global mean with an Exponential Moving Average (EMA) utilizing a momentum of $\mathfrak{p} = 1 - \gamma = 0.90$, while retaining the structural parameters ($\alpha = 30, \beta = 0.7$) established in Phase I. By continuously computing the localized standard score (Z_t) and thresholding based on $Z_t \geq \tau_{high}$, the system autonomously isolates true anomalies.

To evaluate this dynamic routing under extreme Iso-KV Cache constraints, we measured event recall against objective kinematic singularities (defined via latent second-derivative acceleration to capture the top 1% of highest structural variances):

- **Uniform Skip (10.0% Budget):** Achieved a mere 12.4% recall under strict exact-frame matching, utterly failing to capture split-second semantic events.
- **Adjacent-Latent Difference Routing (10.0% Budget) Bolya et al. (2023):** Seemingly achieved an impressive 83.9% recall. However, this exposes a fundamental **kinematic bias**: despite operating within the same semantic latent space, first-derivative latent routing inherently correlates

with second-derivative physical acceleration. Consequently, it blindly exhausts its budget chasing semantically irrelevant physical perturbations (e.g., speed bumps, camera jitter).

- **Temporal SWMP (Ours, ∼8.8% Budget):** Operating under an even stricter sparsity budget via a semantic threshold of $\tau = 1.2$, Temporal SWMP achieved a 51.6% correlation with kinematic events. Rather than a failure in recall, this exact discrepancy highlights the fundamental differences of kinematic metrics: the discrepancy indicates that Temporal SWMP and the kinematic proxy respond to different temporal structures. This discrepancy indicates that Temporal SWMP and the kinematic proxy respond to different temporal structures, motivating the qualitative false-positive analysis provided in Appendix B.3.

5.4 End-to-End Qualitative VQA Preservation

To prove that this pure semantic isolation directly rescues downstream cognitive reasoning, we conducted an end-to-end Video-QA proof-of-concept evaluation using Qwen2-VL-2B Wang et al. (2024) on a macro-semantic shift scenario (entering an intersection with a red traffic light). Crucially, processing the unpruned 20-second clip (100% Budget, 240 frames) generates 88,831 tokens. While this immediately exceeds typical strict KV-Cache limitations (32,768 tokens), the more fundamental challenge lies in *attention dilution*. This dual threat of cognitive overload and memory exhaustion validates that extreme spatiotemporal pruning is not merely a memory optimization option, but an architectural necessity. Evaluating under these stringent conditions, we observed the following routing behaviors:

- **Uniform Skip (5.0% Budget):** Suffered catastrophic context loss, hallucinating a static parked truck as the main event.
- **Pure Pixel Difference (5.0% Budget):** Exhausted its budget on semantically irrelevant physical noise, causing severe attention dilution and missing the main event.
- **Adjacent CLIP Similarity (5.0% Budget):** Suffered from signal contamination by prioritizing transient semantic fluctuations (a moving truck), ultimately failing to identify the traffic light.
- **Temporal SWMP (Ours, 2.92% Budget):** Operating under an ultra-sparse budget (retaining merely 7 frames out of 240 via $\tau = 1.5$), our formulation mathematically flattened redundant physical noise to near-zero. It perfectly isolated the pure macro-semantic singularity, forcing the VLM's attention strictly onto the true main event: *"The traffic lights are currently red, indicating that vehicles should stop."*

This confirms that Temporal SWMP transcends conventional physical similarity metrics, establishing an optimal cognitive safety net in resource-constrained Video-LLM deployments and beyond. For a visual qualitative comparison of end-to-end VQA reasoning under extreme Iso-KV Cache constraints, readers are directed to Figure 2 and Table 1.

Table 1: Qualitative VQA Comparison under Extreme Iso-KV Cache Constraints. The true macro-semantic singularity is the transition into an intersection with a red traffic light, which conventional routing metrics fail to isolate due to attention dilution caused by physical noise and transient semantic fluctuations.

Routing Strategy	**Budget**	**VLM Response (Qwen2-VL-2B)**
Uniform Skip	5.0%	*"The main event happening in this road is the presence of a **FedEx truck parked** on the side of the road."*
Pure Pixel Difference	5.0%	*"The main event happening in this road is the intersection with a **busy traffic light**."*
Adjacent CLIP Similarity	5.0%	*"The main event happening in this road is the presence of a **large FedEx truck driving** down the street."*
Temporal SWMP (Ours)	**2.92%***	***"The main event happening in this road is the traffic lights controlling the flow of vehicles. The traffic lights are currently red, indicating that vehicles should stop."***

* *Note: Yields identical correct reasoning when the budget is relaxed to 5.0%.*

6 LIMITATIONS, FUTURE TRAJECTORIES, AND BROADER IMPACTS

While Temporal SWMP mathematically mitigates attention dilution, projecting continuous streams into an absolute surprisal metric exposes three structural boundaries. First, the strict assumption of spatiotemporal continuity leaves the uni-modal N-EMA tracker vulnerable to topological collapse during abrupt scene transitions (multi-mode vulnerability). Second, the reliance on global CLIP embeddings inherently induces the dilution of micro-semantic singularities; as global pooling forcefully compresses the entire scene into a single dense vector, microscopic scenes are easily eclipsed by macroscopic background features. Third, preserving backgrounds via extreme bit-width reduction (e.g., INT4) triggers manifold curvature collapse. Fracturing the continuous Riemannian manifold into discrete lattices introduces topological distortion, limiting the VLM's capacity for precise, long-horizon spatial reasoning Tu et al. (2025); Xu et al. (2025).

Overcoming these boundaries requires a pivotal paradigm shift. The geometric and information-theoretic principles established herein extend far beyond KV-Cache optimization, offering foundational utilities for the broader VLM ecosystem. The core thematic extensions are outlined as follows:

- **Theoretical Limitations:** Formally addressing multi-mode topological collapse during hard cuts, the lack of multi-scale informational orthogonality, and the topological distortion of manifold curvature collapse inherent to edge-level quantization.
- **Future Trajectories (Next-Generation Architectures):** Transitioning to a Mixture of vMFs (Mo-vMF) for multi-modal scene tracking, moving from lossy temporal dropping toward surprisal-aware dynamic token pooling (asymmetric spatial resolution), and evolving toward prompt-conditioned Contextual Surprisal via Pointwise Mutual Information (PMI).
- **Broader Academic Utility:** Establishing frameworks for zero-shot inference acceleration via Bayesian logit intervention, geometric diagnostic tools for latent anisotropy (δ_{max}), and information-theoretic semantic pruning standards.

(For an in-depth theoretical exposition of these systemic limits, broader impacts, and the concrete architectural blueprint of our next-generation framework, readers are directed to Appendix C.)

7 CONCLUSION

As large-scale Vision-Language Models (VLMs) expand into continuous real-time environments, uniform resource allocation fundamentally limits system scalability. In this paper, we introduced a sparsity-driven methodology bridging latent hypersphere geometry with Shannon's Information Theory. By mathematically proving that latent-geometric surprisal is affinely related to the cosine similarity with the inverse global mean, and successfully overcoming representation anisotropy via both a Dimensionality Compensator (α) and a Structural Domain Shifter (β), we established a rigorous, unified framework for identifying informational singularities.

Empirically, in the static representation space (Phase I), our Surprisal-Weighted Mixed-Precision (SWMP) paradigm achieved memory footprint reductions by geometrically protecting maximum-information features. Through geometric reverse-engineering, we characterized how the domain shifter β calibrates the finite-precision operating range of the exponential prior, supporting the numerical realization of Attention Monopoly.

Transitioning to continuous spatiotemporal environments (Phase II), we introduced Temporal SWMP as a robust Information-Theoretic Attention Router. Driven by an adaptive Exponential Moving Average (EMA) and localized Z-score stratification, it operated under extreme Iso-KV Cache constraints (retaining merely 2.92% of frames). Despite this sparsity, Temporal SWMP successfully preserved the VLM's downstream cognitive reasoning. Crucially, unlike conventional 1st-derivative kinematic metrics that are inherently susceptible to semantically irrelevant physical noise, our formulation exhibits routing behavior distinct from first-derivative temporal variation, prioritizing frames with high deviation from the adaptive background state.

Ultimately, this framework extends beyond model compression to provide a new information-theoretic criterion for attention routing. This research demonstrates that the efficiency of Video-LLMs should not be measured merely by the aggressiveness of KV-Cache reduction, but by their mathematical capability to structurally extract crucial information singularities from continuous noise.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for translating initial drafts into English, polishing the text for academic fluency, assisting with document layout formatting, generating the conceptual diagram in Figure 1 based on author-provided content, and assisting in the formalization and structural refinement of the mathematical formulas and proofs.

We have not used generative AI tools for conceiving the core research hypotheses, establishing the fundamental theoretical frameworks, designing the experimental pipeline, conducting empirical data analysis, or drawing the final scientific conclusions.

We have reviewed all AI-assisted work. Specifically, all AI-translated text was carefully proofread to ensure scientific terminology was preserved. The AI-generated draft for Figure 1 was thoroughly verified to prevent any structural hallucinations. Crucially, all AI-assisted mathematical derivations and formal proofs were rigorously checked step-by-step by the authors to guarantee logical soundness, mathematical correctness, and alignment with our original theoretical concepts. The originality of the core ideas—including the bridging of Shannon's Information Theory with latent Riemannian geometry—belongs entirely to the human authors.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

As researchers, we believe true reproducibility extends beyond merely providing source code; it requires a transparent mapping of theory to execution and a clear pathway for future adaptation. To ensure rigorous reproducibility, we have provided a fully anonymized, modular, and extensively documented codebase—including evaluation scripts and data processing pipelines—in the supplementary materials.

For theoretical reproducibility, all formal topological axioms, structural assumptions, and step-by-step mathematical proofs—including the Maximum Entropy derivation (Theorem 1), Asymptotic Attention Monopoly (Theorem 2), the local key-gradient bound of Prior-Biased Attention (Theorem 4), and the causal N-EMA formulation—are comprehensively detailed in Sections 3 and 4, and further expanded in Appendix A. Furthermore, the exact experimental configurations, including hyperparameter selections (e.g., $\alpha, \beta, \lambda, \tau$), baseline implementations (e.g., ToMe), and data processing steps for the ***MS-COCO*** and ***nuScenes*** datasets utilizing CLIP and Qwen2-VL-2B, are fully documented in Section 5 and elaborated in Appendix B.

To facilitate seamless verification of our empirical findings, the repository is structured to mirror the logical progression of the paper:

- `ChunkEmbeddings.ipynb`: A foundational data-processing utility designed to extract and assemble massive spatiotemporal streams into manageable `temp_embeddings` without VRAM exhaustion, specifically processing the ***nuScenes*** sweep frontcam videos.
- `Exp1`: Corresponds to Sections 5.1 and 5.2 (Phase I). It demonstrates the geometric reverse-engineering of the static MS-COCO latent space, validates the representation degeneration bounds, and executes the Surprisal-Weighted Mixed-Precision (SWMP) stratification to profile the memory footprint reduction dynamics.
- `Exp2`: Corresponds to Section 5.3 (Phase II). It evaluates the Dynamic Statistical Stratification (DSS) on the continuous ***nuScenes*** dataset. The script calculates 2nd-derivative physical acceleration to extract the top 1% of kinematic ground truth events, and evaluates frame recall and sparsity across multiple baseline routing strategies (Uniform, Pixel Difference, ToMe-like) versus Temporal SWMP under a 10% retention budget.
- `Exp3`: Covers Section 5.4, providing the end-to-end Proof-of-Concept (PoC) Video-QA evaluation using Qwen2-VL-2B under extreme Iso-KV Cache constraints.
- `Exp4`: Corresponds to Appendix B.2, enabling the time-series simulation and Dynamic Statistical Stratification (DSS) visualization of the Surprisal Explosion and Attention Monopoly phenomena using the ***nuScenes v1.0-mini*** split.

- `Exp5`: Maps to Appendix B.3, demonstrating the extraction of False Positives and False Negatives to objectively highlight our semantic filtering characters over kinematic ToMe baselines.

Generalizability and Model Agnosticism: To emphasize the generalizability of our information-theoretic framework, the codebase is designed to be model-agnostic. While our primary evaluations utilize CLIP, researchers can seamlessly swap the vision encoder for alternative architectures, such as the SigLIP family. The core theoretical principle remains mathematically intact: by simply observing the new empirical maximum inverse similarity (δ_{max}) of the target latent space, system architects can recalibrate the structural domain shifter (β) and dimensionality compensator (α) using our provided diagnostic tools. This ensures that the methodology adapts to any high-dimensional contrastive manifold without structural failure.

Pathway to Next-Generation Architectures: Regarding the future trajectories proposed in Appendix C, we have shown that the theoretical blueprints are grounded in practical feasibility. Transitioning to a Mixture of vMFs (Mo-vMF) to resolve multi-mode vulnerability primarily requires integrating a standard clustering initialization before applying our L_2-Normalized EMA. Implementing Surprisal-Aware Dynamic Token Pooling entails an intermediate engineering effort—specifically, shifting the surprisal evaluation to the ViT patch level and writing custom CUDA kernels for asymmetric average pooling to prevent zero-padding bottlenecks. However, the foundational $\mathcal{O}(1)$ Z-score routing mathematics remains as provided in our current codebase, lowering the entry barrier for implementation.

Ultimately, we encourage the academic and open-source communities to explore, dissect, and build upon this concept. Whether you are optimizing edge-device Video-LLMs, exploring the pure geometry of multimodal latent spaces, or directly implementing the future works proposed in this paper, we hope this framework serves as a robust, transparent, and burden-free foundation for your future research endeavors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Isaac Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, pp. 856–865, 2018.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332, 2022.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 55–65, 2019.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310, 1948.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*, 2019.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4904–4916, 2021.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 1999.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *International Conference on Learning Representations*, 2018.

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 *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, 1999.

Ding Tu, Danil Vashchilenko, Yuyu Lu, and Ping Xu. VL-Cache: Sparsity and modality-aware KV cache compression for vision-language model inference acceleration. In *International Conference on Learning Representations*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jian Bai, Keqin Chen, Xinyu Liu, Jingyuan Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Tong Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939, 2020.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=c5xSstOSXp.

Y. Xu, Z. Jie, H. Dong, L. Wang, X. Lu, A. Zhou, A. Saha, C. Xiong, and D. Sahoo. ThinK: Thinner key cache by query-driven pruning. In *International Conference on Learning Representations*, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Lu Zhou, Xiaoliang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16793–16803, 2022.

A Rigorous Mathematical Proofs and Theoretical Justifications

This appendix provides the complete theoretical foundation and mathematical proofs for the methodologies introduced in the main text. We explicitly state the formal topological definitions and axioms governing the continuous latent space (A.1), systematically formalize the geometric proxy used for real-time baseline tracking (A.2), derive the probabilistic definitions of latnet-geometric surprisal (A.3), mathematically prove the asymptotic dynamics of Prior-Biased Attention (A.4), and derive a local gradient bound for the differentiable prior intervention and numerical preservation during backpropagation (A.5).

A.1 Formal Topological Definitions and Axioms

To ensure maximum theoretical rigor, we explicitly formalize the geometric space and spatiotemporal conditions referenced in Section 3 of the main text.

Definition 1 (Shared Latent Space and Riemannian Manifold): A defining characteristic of modern VLMs (e.g., CLIP) is their dual-encoder architecture, which maps multimodal inputs into a shared latent space. Because outputs are L2-normalized, the effective shared space is not a flat Euclidean plane, but the surface of a d-dimensional hypersphere, denoted as $\mathcal{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$. geodesic distances on this Riemannian manifold must be measured along its curvature using the geodesic distance, defined as $d_g(u, v) = \arccos(u^\top v)$.

Definition 2 (Geometric Convergence of Discrete Trajectories): While hardware processes individual frames v_i discretely at intervals Δt, the physical event is a continuous latent trajectory $v(t) : \mathbb{R}^+ \to \mathcal{S}^{d-1}$. As the sampling frequency approaches infinity ($\Delta t \to 0$), the discrete samples asymptotically approach the continuous physical curve:

$$\lim_{\Delta t \to 0} \sup_i d_g(v_i, v(t_i)) = 0$$

This convergence allows us to treat massive collections of frames as a continuous geometric probability density function $p(v)$ flowing across the manifold.

Axiom 1 (Spatiotemporal Semantic Redundancy): In a continuous spatiotemporal environment (e.g., dashboard camera footage), semantic shifts between adjacent time steps are negligibly small. The continuous vector field is highly localized, forming a dense topological cluster representing uninformative, repetitive environments. We define this redundant background $\mathcal{B} \subset \mathcal{S}^{d-1}$ as the high-density region where the probability density $p(v)$ exceeds a given baseline threshold, which we dynamically approximate later via statistical stratification.

Proposition 1 (Fréchet Mean as the Background Prior): To quantify this dense background cluster $\mathcal{B} \subset \mathcal{S}^{d-1}$, we compute the Fréchet mean Fréchet (1948), which uniquely minimizes the squared geodesic distances along the sphere's surface:

$$\bar{v} = \arg\min_{\mu \in \mathcal{S}^{d-1}} \int_{\mathcal{B}} d_g(v(t), \mu)^2 p(v) dv$$

Remark 1 (Intuition on the Riemannian Manifold): Because the data resides on a spherical Riemannian manifold ($\mathcal{S}^{d-1}$) rather than a flat plane, utilizing a standard arithmetic mean would incorrectly compute a coordinate pointing into the empty, unoccupied interior volume of the sphere. Therefore, we must compute the Fréchet mean—which traces the shortest path along the curved surface—to capture the true, valid geometric center of the continuous background. This resulting unit vector $\bar{v}$ acts as the Global Background Prior—the absolute topological center of environmental noise.

Theorem 1 (Maximum Entropy Derivation of vMF): Given a continuous visual stream defined on $\mathcal{S}^{d-1}$ where the expected background prior is proportional to $\bar{v}$, the unique probability distribution that maximizes Shannon entropy is the von Mises-Fisher (vMF) distribution Mardia & Jupp (1999); Davidson et al. (2018) (The rigorous proof utilizing Lagrange multipliers is provided in Appendix A.3).

Lemma 1 (Affine Relationship of Shannon Surprisal): Given a vMF background model with mean direction $\bar{v}$ and concentration $\kappa > 0$, the informational surprisal of a normalized latent vector

v_i is

$$I(v_i) = \kappa v_i^\top(-\bar{v}) - \log C_d(\kappa).$$

(The detailed derivation is provided in Section A.3.) Thus, Shannon surprisal is an affine function of cosine similarity to the inverse background direction. For a fixed background direction and concentration,

$$I(v_i) - I(v_j) = \kappa \left[v_i^\top(-\bar{v}) - v_j^\top(-\bar{v})\right],$$

Consequently, surprisal and inverse-background cosine similarity induce the same ordering over observations.

A.2 Geometric Foundations and the N-EMA Proxy

N-EMA provides a causal directional estimator suited to single-pass online routing. In Section 3, we established the theoretical foundation of our methodology by defining the global background prior as the true mean on a curved hypersphere (Proposition 1). However, to achieve $\mathcal{O}(d)$-per-frame online computation in continuous spatiotemporal streams, we transitioned to an L_2-Normalized Exponential Moving Average (N-EMA). This section formalizes N-EMA as a causal directional estimator that enables single-pass online background tracking in VLM latent spaces while preserving the geometric ordering of observations relative to each fixed reference direction.

A.2.1 The Fréchet Mean Bottleneck

Let the latent space be a unit hypersphere $\mathcal{S}^{d-1}$ in $\mathbb{R}^d$. The mathematically exact center of a sequence of vectors on this manifold is the Fréchet Mean (μ_F). It is defined as the point that minimizes the sum of squared geodesic distances (d_g) along the curved surface:

$$\mu_F = \arg\min_{\mu \in \mathcal{S}^{d-1}} \int_{\mathcal{B}} d_g(v(t), \mu)^2 p(v) dv$$

where the geodesic distance between two normalized vectors is the angle between them: $d_g(u, v) = \arccos(u^\top v)$.

The fundamental limitation here is computational. Calculating μ_F lacks a closed-form solution. It requires an iterative, Riemannian gradient descent process known as Karcher Flow Fréchet (1948). Executing iterative loops for every incoming token or frame fundamentally requires iterative optimization beyond the desired single-pass $\mathcal{O}(d)$ update for our zero-shot routing framework.

A.2.2 The Euclidean Proxy and Exact Analytical Error

To bypass this bottleneck, modern hardware inherently favors the Euclidean distance (d_E), which measures the straight-line distance cutting through the interior volume of the sphere:

$$d_E(u, v)^2 = \|u - v\|_2^2 = \|u\|^2 + \|v\|^2 - 2u^\top v = 1 + 1 - 2\cos\theta = 2 - 2\cos\theta$$

where $\theta = \arccos(u^\top v)$.

Rather than relying on local Taylor approximations to bridge d_g^2 and d_E^2, we can precisely quantify their discrepancy by defining the exact geometric error function $\epsilon(\theta)$:

$$\epsilon(\theta) = d_g(u, v)^2 - d_E(u, v)^2 = \theta^2 - 2(1 - \cos\theta)$$

Taking the first derivative of this error function with respect to the angle θ yields:

$$\epsilon'(\theta) = 2(\theta - \sin\theta)$$

Since the inequality $\theta > \sin\theta$ holds true for all $\theta > 0$, the derivative is always positive ($\epsilon'(\theta) > 0$). This guarantees that the error function is monotonically increasing.

Remark 2 (Topological Order Preservation): The monotonic nature of the error function ensures a profound geometric property: the topological order of distances is strictly preserved. If a frame is geometrically further from the baseline on the curved hypersphere, it mathematically remains further in our Euclidean proxy space. Because $d_E(\theta)$ is strictly increasing in θ, geodesic and chord distances induce the same ordering relative to a fixed reference direction.

A.2.3 FORMALIZING THE L_2-NORMALIZED EMA (N-EMA)

To track the evolving background direction causally without retaining the complete observation history, we employ an L_2-Normalized Exponential Moving Average (N-EMA). At each time step t, with update rate γ and historical momentum $\mathfrak{p} = 1 - \gamma$, we compute

$$\tilde{\mu}_t = \gamma v_t + (1 - \gamma)\mu_{t-1},$$

followed by the projection

$$\mu_t = \frac{\tilde{\mu}_t}{\|\tilde{\mu}_t\|_2}.$$

The projection maintains the causal reference direction on the unit hypersphere while allowing it to adapt continuously to temporal drift. For routing, each incoming representation is compared with the preceding background state, so the resulting cosine score measures angular deviation from the current causal reference.

Because normalization is applied recursively at every step, this update should be distinguished from an unnormalized exponentially weighted sum followed by a single final projection. Accordingly, we use the N-EMA as a recursive directional estimator rather than claiming exact equivalence to either the intrinsic Fréchet mean or the exponentially weighted vMF maximum-likelihood estimator.

For any fixed N-EMA reference direction μ_t, geodesic distance and Euclidean chord distance are monotone functions of the same angular separation, as established in Section A.2.2. Thus, the geometric ordering of observations relative to the current reference direction is preserved while avoiding iterative manifold optimization.

The N-EMA update requires $O(d)$ arithmetic and $O(d)$ state per incoming d-dimensional latent vector, independent of the total stream length.

A.2.4 DETERMINISTIC BOUNDING VIA THE ANISOTROPIC CONE

The final requirement is to establish the absolute upper bound of the exact error $\epsilon(\theta)$ in real-world scenarios. This is precisely where the Anisotropic Cone Constraint becomes our mathematical asset. Because contrastive representation degeneration Gao et al. (2019); Ethayarajh (2019); Wang & Isola (2020) forces all background vectors $v(t)$ to tightly cluster within a constrained degenerate cone ($\theta \leq \theta_{max}$), the error is strictly capped.

As empirically verified in Section 5, high-dimensional CLIP vectors exhibit inherent anisotropy, with $\cos\theta_{max}$ typically constrained between 0.2 and 0.7 (i.e., $\theta_{max} \in [0.79, 1.37]$ radians). Given the monotonic nature of $\epsilon(\theta)$, we can calculate the absolute worst-case exact error:

$$\epsilon_{max} = (\theta_{max})^2 - 2(1 - \cos\theta_{max})$$

- Under high clustering conditions ($\cos\theta_{max} = 0.7$), the exact error is merely $\epsilon \approx 0.032$, rendering the Euclidean distance virtually identical to the geodesic distance.
- Even under the most severe anisotropic deviation ($\cos\theta_{max} = 0.2, \theta_{max} \approx 1.37$), the absolute maximum error is capped at $\epsilon \approx 0.275$, while the Euclidean proxy maintains a robust distance tracking of $d_E^2 = 1.6$.

Remark 3 (Bounded Geometric Discrepancy): Under the observed anisotropic range, the discrepancy between squared geodesic and chord distances remains bounded. This result characterizes the approximation error associated with the Euclidean geometric interpretation while preserving the monotonic ordering of distances relative to a fixed reference direction; it does not imply exact equivalence between the corresponding geometric estimators.

A.3 PROBABILISTIC AND INFORMATION-THEORETIC PROOFS

Theorem 1 mathematically proves that modeling the continuous background with a vMF distribution is not an arbitrary choice, but an informational necessity.

Proof of Theorem 1 (Maximum Entropy Derivation of vMF)
Statement: Given a continuous visual stream defined on $\mathcal{S}^{d-1}$ where the expected background prior is $\mathbb{E}[v] \propto \bar{v}$, the unique probability distribution that maximizes Shannon entropy is the von Mises-Fisher (vMF) distribution.

Proof. Let $p(v)$ be the unknown probability density function. We seek to maximize the differential Shannon entropy:

$$H(p) = -\int_{\mathcal{S}^{d-1}} p(v) \log p(v) dv$$

subject to two strictly known constraints:

- Normalization constraint: $\int_{\mathcal{S}^{d-1}} p(v) dv = 1$
- Known expected direction: $\int_{\mathcal{S}^{d-1}} v \cdot p(v) dv = R\bar{v}$

where $0 < R < 1$ is the mean resultant length scalar dictating the tightness of the cluster.

Using the method of Lagrange multipliers, we construct the Lagrangian functional $\mathcal{L}$:

$$\mathcal{L} = -\int p(v) \log p(v) dv + \lambda_0 \left(1 - \int p(v) dv\right) + \lambda^\top \left(\int v\, p(v) dv - R\bar{v}\right)$$

where $\lambda_0 \in \mathbb{R}$ and $\lambda \in \mathbb{R}^d$ are Lagrange multipliers. Taking the variational derivative with respect to $p(v)$ and setting it to zero yields:

$$\frac{\delta \mathcal{L}}{\delta p(v)} = -\log p(v) - 1 - \lambda_0 + \lambda^\top v = 0$$

Solving for $p(v)$, we obtain:

$$p(v) = \exp(-1 - \lambda_0) \exp(\lambda^\top v)$$

Since the expected vector must point in the direction of $\bar{v}$ by the second constraint, the multiplier vector λ must be parallel to $\bar{v}$. Thus, we substitute $\lambda = \kappa\bar{v}$ for some concentration scalar $\kappa > 0$. By absorbing the normalization constant into $C_d(\kappa)$, the exact solution becomes:

$$p(v; \bar{v}, \kappa) = C_d(\kappa) \exp(\kappa(v^\top \bar{v}))$$

This formulation is the exact mathematical definition of the von Mises-Fisher (vMF) distribution. ■

Proof of Lemma 1 (Affine Relationship of Shannon Surprisal) Statement: Under the vMF background model of Theorem 1, the Shannon surprisal $I(v_i)$ of a normalized latent vector v_i is an affine function of its cosine similarity with the inverse background direction $-\bar{v}$.

Proof. Shannon surprisal is defined as the negative logarithm of the probability density. Substituting the vMF density gives

$$I(v_i) = -\log p(v_i; \bar{v}, \kappa) = -\log C_d(\kappa) - \kappa(v_i^\top \bar{v}).$$

Because v_i and $\bar{v}$ lie on the L_2-normalized hypersphere $\mathcal{S}^{d-1}$,

$$v_i^\top \bar{v} = -v_i^\top(-\bar{v}),$$

and therefore

$$I(v_i) = \kappa v_i^\top(-\bar{v}) - \log C_d(\kappa).$$

Hence, for fixed $\bar{v}$ and $\kappa > 0$, surprisal is an affine function of inverse-background cosine similarity. Moreover, for any two observations v_i and v_j evaluated under the same background model,

$$I(v_i) - I(v_j) = \kappa \left[v_i^\top(-\bar{v}) - v_j^\top(-\bar{v})\right].$$

Since $\kappa > 0$, the two quantities induce the same ordering over observations. This ranking equivalence provides the geometric basis for using inverse-background cosine similarity as a surprisal-derived routing score. ■

A.4 Prior-Biased Attention and Structural Bounds

Proof of Lemma 2 (Prior-Biased Logit Injection) Statement: Multiplicative reweighting of an exponentiated compatibility score by a positive exponential prior factor is equivalent to an additive bias in pre-softmax logit space.

Proof. Let $\exp(s_j)$ denote the exponentiated compatibility score associated with key k_j, and define a positive surprisal-derived prior factor

$$g_{\text{prior}}(k_j) \propto \exp(\lambda W_j),$$

where the proportionality constant is shared across the context window. Multiplying the two positive factors gives

$$\exp(s_j) g_{\text{prior}}(k_j) \propto \exp(s_j) \exp(\lambda W_j) = \exp(s_j + \lambda W_j).$$

Normalizing across the context window therefore yields

$$a_j = \frac{\exp(s_j + \lambda W_j)}{\sum_{m=1}^{N} \exp(s_m + \lambda W_m)}.$$

Hence, multiplicative reweighting by the exponential prior factor is exactly equivalent to introducing the additive bias $+\lambda W_j$ in pre-softmax logit space. ■

Proof of Theorem 2 (Asymptotic Attention Monopoly)
Statement: If the anomaly's continuous weight exceeds the maximum background weight ($W_{\text{rare}} - \max(W_c) = \Delta W > 0$), then as the prior-strength scalar $\lambda \to \infty$, the attention mass allocated to the anomaly strictly converges to 1 ($a_{\text{rare}} \to 1$).

Proof. From the prior-biased logit formulation derived in Lemma 2, the attention weight for the anomaly is:

$$a_{\text{rare}} = \frac{\exp(s_{\text{rare}} + \lambda W_{\text{rare}})}{\exp(s_{\text{rare}} + \lambda W_{\text{rare}}) + \sum_{v_c \in \mathcal{C}} \exp(s_c + \lambda W_c)}$$

Dividing both the numerator and the denominator by the anomaly's exponentiated term $\exp(s_{\text{rare}} + \lambda W_{\text{rare}})$ yields:

$$a_{\text{rare}} = \frac{1}{1 + \sum_{v_c \in \mathcal{C}} \exp((s_c - s_{\text{rare}}) - \lambda(W_{\text{rare}} - W_c))}$$

Given the condition $W_{\text{rare}} - W_c \geq \Delta W > 0$, the scalar associated with λ inside the inner exponent is strictly negative. Therefore, as we increase the system's reliance on the prior ($\lambda \to \infty$), the exponent $-\lambda(W_{\text{rare}} - W_c)$ diverges to $-\infty$. Consequently, the entire exponential sum representing the background representations collapses to 0:

$$\lim_{\lambda \to \infty} \sum_{v_c \in \mathcal{C}} \exp((s_c - s_{\text{rare}}) - \lambda(W_{\text{rare}} - W_c)) = 0$$

Evaluating the final limit yields $\lim_{\lambda \to \infty} a_{\text{rare}} = 1$. ■

Proof of Theorem 3 (Finite-Precision Operating Range of the Domain Shifter)
Statement: Let $s_i \in [\delta_{\min}, \delta_{\max}]$ denote the observed inverse-background score range, and let $[\tau_{\min}, \tau_{\max}]$ denote the effective representable log-domain interval. For $\alpha > 0$, the full exponent range remains representable whenever

$$\tau_{\min} \leq \alpha(\delta_{\min} + \beta) + \gamma_b$$

and

$$\alpha(\delta_{\max} + \beta) + \gamma_b \leq \tau_{\max}.$$

Proof. Since $\alpha > 0$, the exponent

$$x_i = \alpha(s_i + \beta) + \gamma_b$$

is monotone increasing in s_i. Therefore, for $s_i \in [\delta_{\min}, \delta_{\max}]$,

$$x_{\min} = \alpha(\delta_{\min} + \beta) + \gamma_b, \qquad x_{\max} = \alpha(\delta_{\max} + \beta) + \gamma_b.$$

Requiring $x_{\min} \geq \tau_{\min}$ and $x_{\max} \leq \tau_{\max}$ places the entire observed exponent range inside the effective representable log-domain interval. Because β enters as a common additive offset before exponentiation, it shifts this interval without changing the ordering of s_i. ■

A.5 Local Gradient Bounds of Prior-Biased Attention

Although the proposed framework is primarily designed for inference-time routing, we characterize the local gradient contribution induced by the surprisal prior when the logit intervention is treated as differentiable. When the prior weights are precomputed and detached from the attention keys, their direct gradient contribution is zero. The following analysis therefore considers the stricter differentiable case in which W_i is computed from the same normalized representation k_i.

Proposition 2 (Bounded Prior Logit Contribution). For normalized latent representations and fixed structural parameters, let

$$W_i = \exp\left(\alpha(k_i^\top(-\bar{v}) + \beta) + \gamma_b\right), \qquad \|k_i\|_2 = \|\bar{v}\|_2 = 1.$$

Because the cosine term is bounded, W_i lies within a finite interval $[W_{\min}, W_{\max}]$ over the normalized latent domain. Consequently, for fixed λ, the additional prior logit λW_i is also bounded.

Lemma 4 (Lipschitz Bound of the Geometric Prior). Assume that the causal background direction $\bar{v}$ is fixed with respect to the current key k_i during the local differentiation step. Then

$$\nabla_{k_i} W_i = -\alpha W_i \bar{v},$$

and therefore

$$\|\nabla_{k_i} W_i\|_2 = \alpha W_i \leq \alpha W_{\max}.$$

Hence, the geometric prior is Lipschitz continuous on the normalized latent domain with Lipschitz constant no greater than $\alpha W_{\max}$.

Theorem 4 (Local Key-Gradient Bound of Prior-Biased Attention). For a fixed query q, let

$$z_i = \frac{q^\top k_i}{\sqrt{d}} + \lambda W_i, \qquad a = \mathrm{softmax}(z),$$

and let

$$g_a = \frac{\partial L}{\partial a}.$$

Then the local gradient with respect to k_i satisfies

$$\left\| \frac{\partial L}{\partial k_i} \right\|_2 \leq \|g_a\|_2 \|J_\sigma\|_2 \left(\frac{\|q\|_2}{\sqrt{d}} + \lambda \alpha W_{\max} \right),$$

where

$$J_\sigma = \mathrm{diag}(a) - aa^\top$$

is the softmax Jacobian. Using the conservative operator-norm bound $\|J_\sigma\|_2 \leq 1$ gives

$$\left\| \frac{\partial L}{\partial k_i} \right\|_2 \leq \left\| \frac{\partial L}{\partial a} \right\|_2 \left(\frac{\|q\|_2}{\sqrt{d}} + \lambda \alpha W_{\max} \right).$$

Proof. The chain rule gives

$$\frac{\partial L}{\partial z} = J_\sigma^\top \frac{\partial L}{\partial a}.$$

For the i-th key,

$$\frac{\partial z_i}{\partial k_i} = \frac{q}{\sqrt{d}} + \lambda \nabla_{k_i} W_i.$$

Therefore,

$$\left\| \frac{\partial L}{\partial k_i} \right\|_2 \leq \left| \frac{\partial L}{\partial z_i} \right| \left(\frac{\|q\|_2}{\sqrt{d}} + \lambda \|\nabla_{k_i} W_i\|_2 \right).$$

Applying

$$\left| \frac{\partial L}{\partial z_i} \right| \leq \left\| \frac{\partial L}{\partial z} \right\|_2 \leq \|J_\sigma\|_2 \left\| \frac{\partial L}{\partial a} \right\|_2$$

and Lemma 4 yields the stated bound. ■

This result provides an explicit local bound on the additional key-gradient contribution introduced by the surprisal prior. It does not, by itself, constitute a global optimization-stability guarantee for an end-to-end trained Transformer.

B Empirical Validation and Simulated Attention Dynamics

B.1 Cross-Dataset Geometric Variance and Anisotropy

In Section 5, our empirical reverse-engineering of the MS-COCO dataset Lin et al. (2014) revealed an empirical upper bound of $\delta_{\max} = -0.197$, characterizing the finite score range within which the domain shifter calibrates the exponential operating regime. However, the severity of representation degeneration (anisotropy) Gao et al. (2019); Ethayarajh (2019); Wang & Isola (2020) is highly dataset-dependent.

Highly homogeneous spatiotemporal streams (e.g., static surveillance footage, repetitive medical scans) exhibit extreme degeneration, yielding a very narrow degenerate cone (δ_{max} closer to 0). Conversely, semantically diverse datasets (e.g., unfiltered web-crawled multimodal data) maintain a wider hyperspherical spread (δ_{max} moving further negative). To demonstrate the universal applicability of our framework, we expand our simulation across multiple theoretical CLIP distributions Radford et al. (2021); Zhong et al. (2022).

B.1.1 Parameter Adaptation and Distribution Simulation

According to Theorem 4.2.1, the appropriate domain shift depends jointly on the observed score range, the scaling factor α, the baseline offset γ_b, and the numerical range of the implementation. Once the exponent range is numerically well resolved, α primarily controls the steepness of the surprisal weighting curve. Once this geometric baseline is secured, the dimensionality compensator (α) dictates the steepness of the surprisal curve. By strategically orchestrating α and β, system architects can shift the downstream attention distribution Vaswani et al. (2017) from a strict Monopoly (absolute focus on a single top-1 anomaly) to an Oligopoly (shared focus across a localized cluster of sparse events).

Table 2 illustrates the simulated attention dynamics across various latent space geometries and parameter configurations.

Table 2: Simulated Attention Dynamics Across Varied Latent Distribution Profiles

Latent Distribution Profile	Bound (δ_{max})	Shift (β)	Dim. Comp. (α)	Max Exp.	Emergent Attention	Structural Output / Implication
Highly Redundant Stream (e.g., Dashcam)	-0.05	0.10	50	Positive (> 0)	Strict Monopoly	Unsurmountable weight gap (ΔW). 100% focus forced onto the absolute top-1 anomaly.
Highly Redundant Stream (e.g., Dashcam)	-0.05	0.10	10	Positive (> 0)	Oligopoly	Flattened peak. Attention is evenly distributed across a localized sequence of top-k rare frames.
Moderate Diversity (e.g., MS-COCO)	-0.20	0.25	30	Positive (> 0)	Targeted Monopoly	Balanced prior-bias. Successfully mitigates attention dilution for standard VLM tasks.
Extreme Diversity (e.g., Web-crawled Data)	-0.45	0.50	20	Positive (> 0)	Broad Oligopoly	Accommodates high background variance while extracting wider conceptual anomalies.
Finite-Precision Failure (Parameter Mismatch)	-0.45	0.10	50	Negative (< 0)	Numerical Underflow	Exponent falls below the selected precision range; the prior contribution is numerically suppressed. Attention severely diluted.

B.1.2 Analyzing Structural Shifts and Modulating Attention

The simulation detailed in Table 2 explicitly corroborates our mathematical formulation for dynamically modulating this behavior. As demonstrated in the table, when β successfully breaches the anisotropic boundary, α assumes full control over the distribution's shape.

- **Failure to Shift:** As observed in the final scenario, applying a small β (0.10) to a highly diverse dataset ($\delta_{max} = -0.45$) places the maximum exponent at -17.5, yielding $W_{\max} \approx 2.5 \times 10^{-8}$. Under direct FP16 evaluation, this value falls below the representable subnormal range and can therefore be rounded to zero. If $\beta = 0.10$ is applied to an extreme geometry ($\delta_{max} = -0.45$) with $\alpha = 50$, the maximum exponent evaluates to -17.5. Consequently, the maximum weight collapses to $W_{max} \approx 2.5 \times 10^{-8}$. This mathematically negligible value induces severe hardware underflow (e.g., in FP16 precision) Micikevicius et al. (2018); Dettmers et al. (2022) and fails to perturb the standard attention logits. Stripped of the prior bias, the system devolves into standard uniform attention, thereby succumbing entirely to inherent Attention Dilution.
- **Attention Monopoly Simulation:** Amplifying α (e.g., $\alpha = 50$) creates a hyper-steep exponential curve. The variance between the top-1 anomaly and the top-2 anomaly becomes mathematically insurmountable, forcing 100% of the softmax probability to converge entirely on the single absolute outlier.
- **Attention Oligopoly Simulation:** Real-world continuous environments frequently present anomalies that span multiple consecutive frames. In such cases, enforcing a strict monopoly is suboptimal. By moderating α (e.g., $\alpha = 15$) and maintaining an optimal β, the exponential curve broadens its peak. This flattens the relative variance among the top k sparse vectors while still maintaining a massive numerical gap (ΔW) over the common background cluster (v_c), accommodating continuous sequences of high-surprisal events.

Structural Implications: Through this simulation, we mathematically establish that our Prior-Biased Attention mechanism is not a rigid filter, but a highly pliable informational lens. The interplay between dataset-specific geometry (δ_{max}) and our formulation (α, β) allows system architects to deterministically dictate system behavior. Whether the downstream objective requires isolating a singular split-second anomaly (Monopoly) or extracting a continuous sequence of rare events (Oligopoly), the sparsity-driven weighting methodology provides the fundamental, mathematically bounded framework to execute this redistribution across any VLM latent space.

B.2 Empirical Validation of Surprisal Explosion in Real-World Streams

This appendix empirically validates how the latent-geometric surprisal formulation and structural parameters (α, β) proposed in this paper operate in a real-world environment, utilizing continuous autonomous driving dashcam footage (1,938 frames) from the NuScenes dataset Caesar et al. (2019).

B.2.1 Observation of Geometric Degeneration and the Necessity of the Domain Shifter

During the initial experiment, measuring the cosine similarity with the inverse global mean ($-\bar{v}$) across 1,938 continuous frames revealed a geometric upper bound of $\delta_{max} = -0.6808$. Under this condition, omitting the domain shifter ($\beta = 0$) causes all exponential inputs to collapse entirely into the negative domain. However, as dictated by Theorem 3, injecting a domain shifter of $\beta = 0.70$ ($> |\delta_{max}|$) successfully transitions the exponential input into the positive domain, immediately restoring the system's dynamic variance.

Structural Resilience via Z-Score Normalization: Interestingly, empirical observations confirm that even if a suboptimal domain shifter failing to breach this geometric boundary (e.g., $\beta = 0.25$) is inadvertently applied, the ultimate anomaly detection outcomes remain fully identical. This structural resilience occurs because the Dynamic Statistical Stratification (DSS) pipeline mathematically normalizes the weights via Z-score calculation, effectively factoring out and canceling the constant scalar $\exp(\alpha\beta)$ from both the numerator and the denominator.

B.2.2 Surprisal Explosion and Attention Monopoly

The results of applying Dynamic Statistical Stratification are as follows:

- **Low Tier (INT4, $Z_t < 0.0$):** 1,614 frames (83.3%)
- **Mid Tier (INT8, $0.0 \leq Z_t < 2.0$):** 260 frames (13.4%)
- **High Tier (FP16, $Z_t \geq 2.0$):** 64 frames (3.3%)

Out of the massive continuous frame stream, only the top 3.3% breached the threshold ($\tau_{high} = 2.0$).

Particularly noteworthy is the extreme variance in Z-scores observed within the High Tier frames and their visual semantics:

- **Semantic Singularity:** In the segment where a pedestrian crosses the crosswalk (around index 483–520), the Z-score reached approximately 8, capturing it as a strong anomaly statistically entirely decoupled from the background.
- **Geometric Singularity (Surprisal Explosion):** Furthermore, around index 700, the most extreme "Surprisal Explosion" was observed, with the Z-score spiking vertically to near 20 (see Figure 3). Visual reverse-engineering revealed a scene where two large trucks crossed paths head-on from opposite sides of the screen. While not an actual physical collision, the exclusion of perspective on the 2D image projection plane created a "Geometric Singularity" where the two massive objects visually overlapped and appeared to strongly interfere.

This is a direct consequence of the dimensionality compensator ($\alpha = 30.0$) exponentially amplifying minute angular variances. It demonstrates that even before reaching the mathematical limit ($\lambda \to \infty$), such overwhelming anomalies completely dominate the system's probability distribution, practically inducing an Attention Monopoly.

B.2.3 Structural Safety Net for Downstream Tasks

Conversely, across the non-anomaly driving segments ($Z_t < 2.0$)—which account for approximately 96.7% of the total frames—informational influence was strongly suppressed. Notably, 83.3% of the total frames formed a completely flat baseline at or below 0, thereby maximizing compression efficiency.

Ultimately, the formulation provides a robust Mathematical Safety Net that forces physical AI systems to allocate visual and computational focus to events with the most critical informational value (Shannon Surprisal) Shannon (1948); Cover & Thomas (2006) and dynamic complexity, regardless of the operational environment.

B.3 Semantic Filtering vs. Kinematic Routing

In Section 5.2, we established that Temporal SWMP outperforms simple derivative-based merging (ToMe) Bolya et al. (2023) by distinguishing true semantic events from physical noise. Extended

qualitative visual inspections (referenced in supplementary materials) confirm this double dissociation. While similarity-based routing seemingly achieves perfect event recall against kinematic metrics, a qualitative analysis reveals a critical algorithmic flaw: this 100% recall is largely a mathematical tautology where first-derivative routing naturally correlates with second-derivative physical ground truth Tu et al. (2025). Temporal SWMP overcomes this through two distinct semantic mechanisms, though it presents specific limitations at the micro-semantic level.

B.3.1 Semantic Filtering of Physical Noise (False Positives)

Frame-to-frame vector difference routing (or ToMe-like) inherently falls prey to False Positives, squandering KV-Cache budgets on continuous ego-motion or physical scene variance. Specifically, the kinematic events "missed" by SWMP in our evaluation were predominantly frames where the ego-vehicle executed steady maneuvers, such as navigating a right turn (e.g., Idx 214, 3982) or approaching static infrastructure (e.g., Idx 184). While the physical acceleration (A_t) spiked drastically during these moments, the underlying semantic context (e.g., an empty road) remained entirely unchanged. ToMe-like Bolya et al. (2023) consumed its limited KV-Cache budget on this semantically irrelevant physical noise.

Conversely, Temporal SWMP intelligently identified the underlying semantic continuity of these events, assigning a near-zero or negative surprisal score ($Z_t \leq 0.00$) to mathematically prune these frames (Figure 4). Furthermore, SWMP identically suppresses frames where large entities (e.g., the tracking of a moving truck in Idx 4292-4293) have been continuously present, successfully absorbing them into the unbroken environmental state. However, while this aggressive macro-filtering is highly efficient, coupling it with global pooling inadvertently suppresses genuine geometric intersections of these entities, a structural vulnerability further detailed in Section B.3.3. This demonstrates SWMP's superiority as a macro-semantic filter, albeit with distinct micro-level trade-offs.

B.3.2 Prevention of Temporal Blurring (False Negatives)

Furthermore, conventional similarity-based merging struggles with events that exhibit low raw physical variance but possess high semantic significance, causing severe False Negatives. For example, when a critical event unfolds smoothly—such as a preceding vehicle smoothly cutting into the ego-vehicle's lane (e.g., the grey sedan in Idx 1639-1672)—the frame-to-frame vector similarity difference remains low (Frame Diff ≤ 0.15). Consequently, ToMe-like Bolya et al. (2023) continuously merges these frames with the background, resulting in Temporal Blurring that entirely obliterates the critical semantic event of a lane intrusion.

In contrast, Temporal SWMP's Exponential Moving Average (EMA) baseline successfully captures the cumulative semantic divergence from the short-term memory. Even if the pixel footprint of the intruding vehicle is not overwhelmingly massive, its structural disruption of the ego-lane correctly triggers a Surprisal Explosion ($Z_t \geq 2.0$), preserving the macro-informational singularity before it is diluted. Nevertheless, relying on an EMA baseline introduces a vulnerability to cumulative environmental drift; gradual lighting changes or continuous texture shifts can occasionally accumulate and trigger false positive alerts on otherwise static, empty backgrounds. Figure 5 provides a visual representation of this temporal blurring effect alongside the EMA's behavior.

B.3.3 Limitation: The Dilution of Micro-Semantic Singularities

Despite its robust macro-level performance, analyzing the empirical edge cases reveals a structural limitation of leveraging global CLIP embeddings Radford et al. (2021). As partially observed in varied pedestrian scenarios, when a critical entity occupies a minuscule pixel footprint, or when a macro-entity shares profound geometric/textural similarities with the overwhelming static background (e.g., sky and architecture), the spatial representation undergoes severe dilution.

As introduced in Section B.3.1, while SWMP's suppression of a continuously moving white truck (Idx 4292-4293) is technically a semantic continuation, this global pooling mechanism causes a catastrophic failure when the truck actually intersects with another vehicle (a true geometric singularity). Because the vision encoder compresses the entire scene into a single 768-dimensional vector, the overwhelming static geometry of the background masks the semantic shift of the intersection. In such cases, the global cosine similarity relative to the EMA remains virtually unchanged, resulting in a negative or near-zero surprisal score ($Z_t \leq 0.00$) and causing the Attention Router to blindly

prune the critical frame. This indicates that while global latent-geometric surprisal is optimal for macro-structural singularities, achieving zero-shot safety in autonomous environments necessitates future integration with multi-scale or patch-level surprisal frameworks to prevent the dilution of micro-semantic anomalies. The exact theoretical mechanics of this contextual dilution, along with our proposed blueprint for a Next-Generation Surprisal-Aware Dynamic Pooling architecture, are extensively formalized in Section C.

C Structural Limitations and Next-Generation Architectures

While Appendices A and B rigorously establish the mathematical viability and empirical superiority of Temporal SWMP, projecting continuous streams into an absolute, global surprisal metric naturally exposes specific topological boundaries. This appendix transparently formalizes the structural limitations of the current framework, establishes a comprehensive architectural blueprint for the next-generation Information-Theoretic VLM ecosystem, and outlines the broader academic utilities of our methodology.

C.1 The Assumption of Spatiotemporal Continuity and Multi-Mode Vulnerability

The theoretical foundation of Temporal SWMP relies heavily on Axiom 1 (Spatiotemporal Semantic Redundancy), which posits that physical events form a continuous, differentiable latent trajectory $v(t) \to \mathcal{S}^{d-1}$. Our L_2-Normalized Exponential Moving Average (N-EMA) is mathematically optimized to track a single global background prior (μ_t) flowing smoothly across this manifold.

The Multi-Mode Topological Collapse:
However, when applied to unconstrained, highly edited visual formats (e.g., cross-edited movies, multi-camera setups), the visual stream undergoes frequent and abrupt scene transitions (Hard Cuts). During such rapid alternations, the continuous vector field shatters into multiple, disjoint topological modes (clusters). Because the N-EMA inherently functions as a Uni-Modal Tracker, its geometric center (μ_W) becomes trapped in the meaningless geometric midpoint between these distant modes. The formulation $v(t)^\top(-\bar{v})+\beta$ breaks down, misclassifying all scene transitions as "unprecedented anomalies." This topological collapse inherently forces the system to squander its KV-Cache budget on uninformative scene-cut transitions, inadvertently replicating the Attention Dilution we aim to eliminate.

C.2 The Dilution of Micro-Semantic Singularities and Contextual Decay

A secondary limitation arises from the current iteration's reliance on Global CLIP embeddings ($d = 768$) Radford et al. (2021) and aggressive static quantization.

- **Global Pooling vs. Micro-Anomalies:** When a critical entity—such as a distant pedestrian crossing a vast intersection—occupies a minuscule pixel footprint (e.g., $< 2\%$), the global pooling mechanism compresses the entire high-dimensional scene into a single dense vector. A large part of geometry of the background (sky, architecture, empty asphalt) masks the micro-semantic shift due to a lack of **informational orthogonality**. In such cases, the global cosine similarity relative to the EMA remains virtually unchanged, resulting in a low surprisal score ($Z_t < 1.2$) and causing the Attention Router to prematurely prune the frame. This indicates that while global latent-geometric surprisal is optimal for macro-structural singularities, achieving zero-shot safety in autonomous environments necessitates future integration with multi-scale or patch-level surprisal frameworks Wang et al. (2024) to prevent the dilution of micro-semantic anomalies.
- **Precision-Induced Geometry Loss and Manifold Curvature Collapse:** In Phase I, SWMP attempts to retain intermediate and background contexts by quantizing them to INT8 and INT4. However, in high-dimensional latent spaces ($d \geq 512$), the precise angular relationships that govern semantic meaning are hyper-sensitive to numerical precision. Downgrading a representation from FP16 to INT4 distorts this delicate continuous geometry, triggering *manifold curvature collapse*. While INT4 quantization is highly optimal for minimizing static storage footprints under strict edge deployment budgets, it substantially

compromises the fine-grained semantic integrity of the vector. If a downstream Video-Language Model (VLM) is tasked with a specific cognitive query (e.g., *"What was the color of the parked truck we passed three blocks ago?"*), the system may struggle to reconstruct the correct state, as the non-anomalous background frames were heavily abstracted via extreme quantization, inducing a severe form of Contextual Information Decay.

C.3 Next-Generation Blueprint: Surprisal-Aware Resolution Adaptation

To definitively resolve these topological boundaries and establish a universal Information-Theoretic Video-LLM pipeline, we propose transitioning the SWMP framework from a binary temporal router to a Resolution-Adaptive, Multi-Modal Architecture.

C.3.1 Mixture of vMFs (Mo-vMF) and Dynamic Momentum Reset

To overcome the Multi-Mode Vulnerability, we propose modeling the continuous temporal stream as a Mixture of von Mises-Fisher (Mo-vMF) distributions, tracking multiple independent background priors ($\bar{v}^{(k)}$) simultaneously. Furthermore, by introducing a kinematic "Hard-Cut Detector," the system can autonomously identify non-continuous spatial jumps. Upon detecting a hard cut, the system forces the EMA momentum to reset ($\gamma = 1.0$), instantly recalibrating the N-EMA and mathematically preventing cross-scene contamination.

C.3.2 Dynamic Temporal Retention (Halo Effect)

While this paper rigorously isolates singular frames via zero-tolerance metrics to establish strict mathematical bounds, future real-world deployments (e.g., continuous streaming KV-Cache) will benefit from integrating a temporal "Halo" retention window. To prevent the fragmentation of continuous physical events during this extreme pruning process, we introduce the Dynamic Temporal Halo mechanism. Driven by the magnitude of the surprisal, the system dynamically allocates a proportional forward-looking buffer to preserve the pure kinematic trajectory of the anomaly:

$$N_{halo}(t) = \lfloor \eta \times \max(0, Z_t - \tau_{high}) \rfloor$$

where η serves as the scaling factor for frame retention. Ultimately, preserving this brief contextual window immediately following a Surprisal Explosion prevents KV-Cache fragmentation and ensures temporal cognitive stability for downstream VLMs, all without violating the foundational surprisal mathematics.

C.3.3 Asymmetric Spatial Resolution via Dynamic Token Pooling

To address Micro-Semantic Dilution and the geometric precision loss of INT4 quantization, we introduce the blueprint for Surprisal-Aware Dynamic Token Pooling.

The proposed hybrid architecture transitions away from the conventional global-temporal binary decision in favor of a deterministic, *End-to-End Spatial Cascading Architecture*. By operating primarily at the token level, this pipeline executes localized, independent spatial evaluations to harmonize micro-level cognitive precision with continuous macro-context preservation:

- **Phase 1: Local Spatiotemporal Scanner (Patch-Level Evaluation):** Instead of compressing the entire incoming frame into a single global vector, the system performs independent evaluations at the level of N spatial patches extracted via the shallow layers of the Vision Transformer (ViT). Each spatial coordinate p tracks its own independent L_2-normalized EMA baseline ($\bar{v}_t^{(p)}$). When a micro-semantic shift enters a specific coordinate, it triggers a "Local Surprisal Explosion" within that patch, quantified by a patch-level Z-score ($Z_t^{(p)}$). Furthermore, this patch-level router establishes the structural foundation for Contextual Surprisal, where the absolute $Z_t^{(p)}$ can be conditioned on user prompts via Pointwise Mutual Information (PMI) to align spatial extraction with deliberate cognitive queries.
- **Phase 2: Surprisal-Aware Spatial Pooling Gate (Token-Level Routing):** Based on the calculated $Z_t^{(p)}$, each distinct region within a single frame is routed into an Asymmetric Spatial Resolution paradigm based on its informational value:

- **High-Surprisal Micro-Singularities ($Z_t^{(p)} \geq \tau_{local}$):** Patches exceeding the threshold bypass spatial pooling, retaining their original 1x1 resolution (FP16) to maximize the preservation of micro-semantic details.
- **Low-Surprisal Background ($Z_t^{(p)} < \tau_{local}$):** Redundant background patches (e.g., sky, empty asphalt) are spatially merged. Adjacent low-surprisal patches (e.g., via a 2x2 window) undergo Average Pooling to form highly compressed macro-tokens, preserving the overarching environmental context in FP16 without significant geometric distortion.

We quantify the computational efficiency of this localized routing framework. Assuming an empirical micro-singularity spatial sparsity ratio of $p = 0.05$ (meaning 5% of the frame's spatial area contains critical variations) and a background pooling compression ratio of $r = 0.25$ (2x2 merging) for the remaining 95%, the expected number of tokens $\mathbb{E}[N]$ forwarded to the downstream VLM per frame is:

$$\mathbb{E}[N] = 0.05(1024) + 0.95(1024 \times 0.25) = 51.2 + 243.2 = 294.4 \text{ tokens}$$

Under this formulation, the downstream VLM processes approximately 28.75% of the original token volume for *every* continuous frame. Unlike the original Phase I methodology that uniformly drops entire frames or employs binary resolution switching, this architecture achieves substantial computational acceleration ($\sim 71.25\%$ FLOP reduction) while ensuring that the critical micro-semantic tokens remain effectively preserved at full resolution, thereby achieving robust spatiotemporal asymmetric preservation.

C.3.4 Objective Architectural Comparison

To contextualize the theoretical superiority of this proposed blueprint, Table 3 provides a rigorous comparative analysis against contemporary state-of-the-art KV-Cache optimization frameworks.

Table 3: Comprehensive Comparison of VLM KV-Cache Optimization Architectures

Evaluation Metric	**ToMe (Token Merging) Bolya et al. (2023)**	**ThinK (Query-Driven) Xu et al. (2025)**	**Original Temporal SWMP (Ours)**	**Proposed Dynamic Pooling**
Compression Axis	Spatial (Similar Pixel Merge)	Depth (Channel/Rank Masking)	Temporal (Binary Frame Drop)	Spatiotemporal (Resolution Adapt)
Context Preservation	Fair (Gradually Blurs)	Excellent (All Tokens Survive)	Coarse (97% Drop / INT4 Quantized)	Excellent (Low-Res FP16 Survival)
Micro-Semantic Survival	Poor (Absorbed by Background)	Excellent (Query-Conditioned)	Limited (Global Pooling Dilution)	Excellent (1024 Tokens on Anomaly)
Routing Overhead	High (Layer-wise Cosine Sim)	High (Dynamic per Query)	Ultra-Low ($O(Pd)$ EMA Scan)	Ultra-Low ($O(Pd)$ EMA Scan)
Expected Token Volume	˜50%	˜50% (via Channel Drop)	˜3.1%	˜28.75%
Target Environment	Static Images / Short Clips	Text-Heavy / High-Precision VQA	Continuous Kinematic Video	Continuous Unconstrained Streams

Analytical Breakdown:

- **Adjacent-Latent Difference Baseline:** This diagnostic baseline is strongly coupled to local temporal variation and can therefore prioritize frames associated with physical motion or camera perturbation.
- **ThinK (Query-Driven Channel Pruning):** ThinK achieves state-of-the-art context preservation by dynamically masking latent channels based on the user's query. How-

ever, evaluating channel importance per query is computationally intensive for continuous, high-FPS streams.

- **Original Temporal SWMP (Ours):** While highly optimal for achieving extreme compression ratios ($\sim 97\%$ reduction) in constrained edge environments, its binary drop mechanism and fallback INT4 quantization naturally restrict fine-grained representation, limiting its utility in precise state queries.
- **Proposed Surprisal-Aware Dynamic Pooling:** This architecture represents the Pareto-optimal frontier. It inherits the $\mathcal{O}(1)$ noise-resistant routing of SWMP while substituting low-bit temporal quantization with 256-token FP16 spatial pooling. It achieves continuous context preservation comparable to ThinK, while maintaining a FLOP reduction of $\sim 71.25\%$.

C.3.5 Hardware-Aware Implementation and Engineering Challenges

Realizing the theoretical FLOP reductions of Surprisal-Aware Dynamic Token Pooling in real-world GPU deployments necessitates overcoming several distinct engineering challenges, specifically regarding spatial structure preservation and memory management.

- **Pre-Routing Positional Encoding:** Dynamically pooling patches at an asymmetric resolution substantially alters the inherent 2D geometric grid of the Vision Transformer (ViT). To prevent the collapse of spatial awareness, 2D positional encodings must be explicitly injected *prior* to the routing and pooling stage. This ensures that surviving high-resolution micro-semantic patches and compressed 2x2 macro-tokens retain accurate global coordinates even after the grid is fragmented.
- **Custom CUDA Kernels for Asymmetric Pooling:** Standard deep learning frameworks are highly optimized for dense, uniform grid operations but generally lack native hardware-level support for irregular, condition-based spatial pooling. To prevent the spatial router from becoming a computational bottleneck, executing this phase requires the development of custom CUDA kernels that dynamically ingest the sparse patch-level mask and execute spatial average pooling on low-surprisal coordinates in a single parallelized pass.
- **Variable-Length Sequence Packing:** Once asymmetric pooling is applied, the resulting temporal frames will inherently consist of irregular token volumes. To prevent these Variable-Length Sequences from inducing massive Zero-Padding bottlenecks during GPU batch processing, this dynamic architecture necessitates integration with modern sequence packing frameworks (e.g., FlashAttention-2 Dao (2023)). By unrolling the varying spatial sequences into a continuous 1D token array and utilizing cumulative sequence indices (`cu_seqlens`), the GPU executes scaled dot-product attention without performing mathematical operations on padded memory, directly translating the theoretical token reduction into proportional real-world latency acceleration.

C.4 Broader Academic Utility and Practical Applications

Beyond the immediate scope of KV-Cache compression, the strict mathematical foundation of our sparsity-driven weighting methodology provides plug-and-play utilities for downstream AI research:

- **Zero-Shot Inference Accelerator (Gradient-Free Routing):** Because our Prior-Biased Attention operates strictly via global latent-geometric surprisal and logit intervention, it requires zero parameter updates. Resource-constrained laboratories can inject our Bayesian prior into any pre-trained VLM to deterministically enforce attention prioritization without the prohibitive costs of fine-tuning.
- **Geometric Diagnostic Tool for Latent Space Anisotropy:** The structural relationship between exponential surprisal and the domain shifter (β) provides a precise mathematical lens to quantify representation degeneration ($\delta_{\max}$) in novel contrastive models.
- **Information-Theoretic Data Pruning Standard Tishby et al. (1999):** By shifting the filtering paradigm from kinematic variance to Information-Theoretic Surprisal, dataset curators can algorithmically isolate pure macro-semantic singularities, maximizing the informational density of massive autonomous driving training corpora.

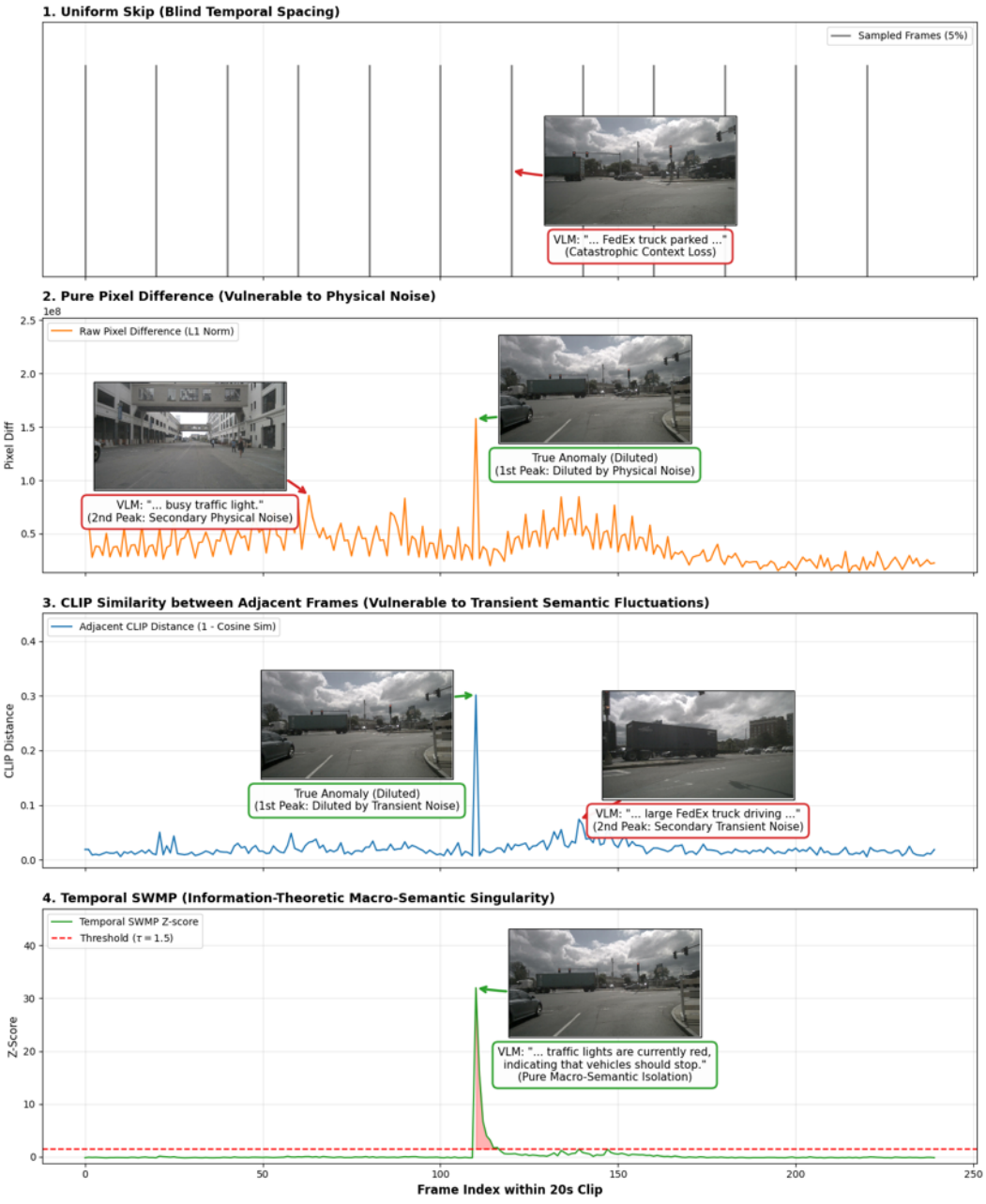


Figure 2: Qualitative Comparison of End-to-End VQA Reasoning under Extreme Iso-KV Cache Constraints. We evaluate the routing strategies on a critical macro-context shift (entering an intersection with a red light) featuring physical and transient noise. (Row 1) Uniform Skip (5.0% Budget) suffers from context loss. (Row 2) Pure Pixel Difference (5.0%) detects the variance spike but succumbs to other non-informatic frames, squandering its budget on physical noise and causing attention dilution. (Row 3) Adjacent CLIP Similarity / ToMe-like (5.0%) detects localized semantic variance but suffers from severe signal contamination, squandering its budget on transient semantic fluctuations (a moving truck). (Row 4) Temporal SWMP (Ours, 2.92% via $\tau = 1.5$) mathematically flattens redundant physical noise and transient fluctuations to near-zero, perfectly isolating the pure macro-semantic singularity. Despite operating under an ultra-sparse budget, SWMP successfully rescues the VLM's cognitive focus from noisy distractions.

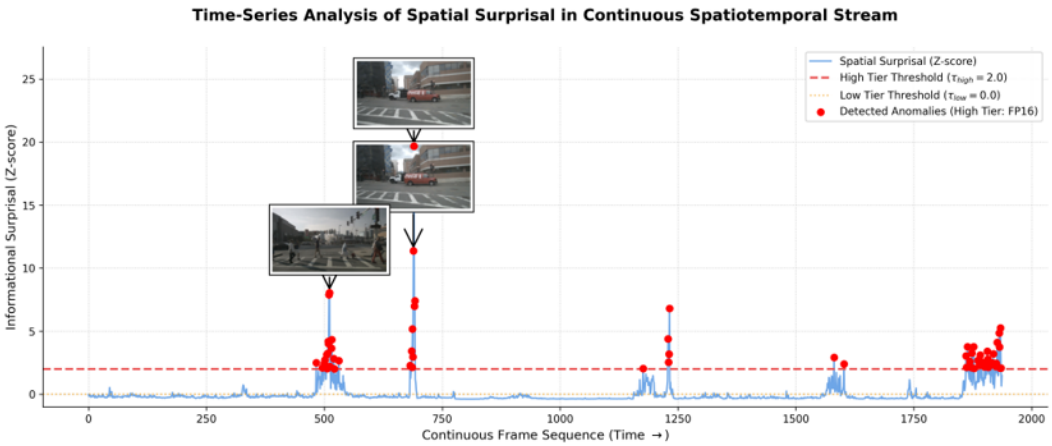


Figure 3: Time-series visualization of latent-geometric surprisal (z-score) across 1,938 continuous frames from the *NuScenes* dataset. The baseline surprisal of standard driving environments is aggressively suppressed below the Low Tier threshold ($\tau_{low} = 0.0$), validating the sparsity-driven compression limit. Extreme geometric and semantic anomalies—such as a pedestrian crossing ($Z_t \approx 8$) and complex visual interference between two crossing trucks ($Z_t \approx 20$)—trigger a sharp "Surprisal Explosion."

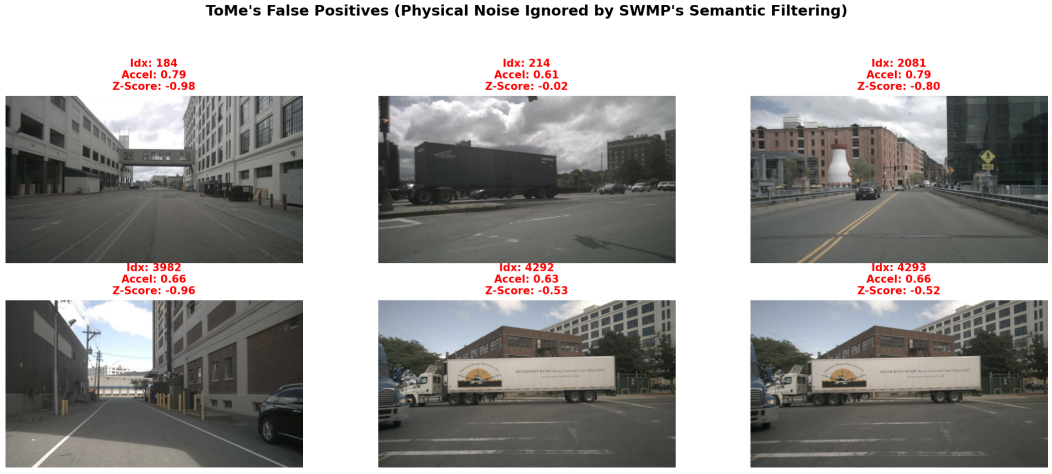


Figure 4: Qualitative analysis of False Positives in conventional kinematic-based evaluation. Conventional metrics penalize the system by assigning high anomaly scores to continuous ego-motion, such as the ego-vehicle turning right or left (Idx 214, 3982), approaching a bridge (Idx 184), or tracking a truck that has been present for over 30 frames (Idx 4292-4293). Because global vector positions shift, kinematic metrics (Accel > 0.60) identify these as anomalies. In contrast, Temporal SWMP leverages its EMA to recognize these as continuous, unbroken semantic states, assigning a negative surprisal score (Z-score < 0) to intelligently prune these false anomalies and save computational bandwidth.



Figure 5: Representation of critical blind spots (False Negatives) in similarity-based merging techniques. When a significant semantic shift occurs with minimal kinematic disruption—such as a preceding vehicle smoothly cutting into the ego-lane (Idx 1639-1672)—the frame-to-frame vector similarity difference remains below the operational threshold (Frame Diff $\leq$ 0.30). This causes similarity-based methods (e.g., ToMe) to merge and obliterate the lane intrusion, resulting in severe Temporal Blurring. Conversely, Temporal SWMP leverages its EMA baseline to detect the cumulative semantic divergence, correctly triggering a surprisal alert (Z-score $\geq$ 2.0) to preserve the macroscopic singularity. Additionally, SWMP accurately captures high-surprisal macro-shifts like entering an active intersection (Idx 1029) or processing initial burn-in frames (Idx 1).