SPECTRA: Spectral Isotropy-Guided Training-Free Temporal Intervention for Long-Video VLMs

Anonymous ECCV 2026 Submission

Paper ID #*****

Abstract. Long-video vision-language models often fail to use distant evidence because attention becomes highly concentrated on a few keys. We revisit this failure from a second-order viewpoint and show that temporal mRoPE anisotropy is controlled by two coupled conditions: block-wise isotropy within each rotary pair and cross-block decoupling across different frequencies. We further prove that, under finite discrete support, phase cancellation is incomplete for near-frequency pairs, leaving residual cross-subspace coupling that sharpens the covariance spectrum and compresses attention coverage. Motivated by this mechanism, we propose **SPECTRA**, a training-free temporal-only intervention. SPECTRA estimates per-head degradation through effective rank, allocates intervention strength with layer-head dual gating, and injects controlled Gaussian interpolation into temporal Q/K channels. The update preserves architecture, requires no retraining, and follows directly from the derived covariance dynamics.

Keywords: Long-Video VLM · mRoPE · Attention Anisotropy · Effective Rank · Training-Free Inference

1 Introduction

Long-video understanding requires integrating evidence across hundreds or thousands of frames. In mRoPE-based VLMs, a recurring failure mode is attention concentration: a few keys absorb most mass, and distant evidence receives little probability. This directly hurts temporal reasoning and multi-event integration.

Existing long-context fixes mostly adjust positional scaling or interpolation. They improve robustness, but they are mainly design heuristics and do not identify the structural source of collapse at layer/head level. Our goal is to derive a clear mechanism and translate it into a training-free intervention.

We show that long-video collapse is fundamentally a second-order problem in temporal channels. The key chain is

$$\begin{aligned} \text{incomplete phase cancellation} &\Rightarrow \text{cross-block covariance coupling} \\ &\Rightarrow \text{spectral peakedness} \\ &\Rightarrow \text{coverage compression.} \end{aligned} \tag{1}$$

The first implication comes from finite-support mRoPE phase analysis; the second and third come from covariance decomposition and logit-variance bounds.

Guided by this chain, we propose **SPECTRA**, a training-free temporal intervention. SPECTRA estimates degradation by effective rank, allocates strength with layer-head dual gating, and applies controlled Gaussian interpolation only to temporal Q/K channels on valid video tokens. The update is plug-and-play and architecture-preserving.

Contributions.

1. We provide a block-structured decomposition showing that global isotropy requires both intra-block isotropy and inter-block decoupling.
2. We analyze discrete phase cancellation and identify finite-support near-frequency coupling as a key source of residual anisotropy.
3. We prove a spectral-to-coverage link: spectral peakedness increases logit extremeness and compresses effective attention span.
4. We propose SPECTRA, a theory-aligned, training-free temporal intervention with explicit covariance-level effects and low overhead.

2 Related Work

2.1 RoPE and Long-Context Extrapolation

Long-context extrapolation for RoPE typically uses scaling, interpolation, or frequency remapping. These approaches improve stability for long sequences, but most analyses are frequency-local. Our perspective is complementary: we study the global second-order matrix after position aggregation and show that local smoothing is insufficient when cross-frequency blocks remain coupled. Boundary of our contribution: we do not propose another frequency schedule; we provide a structural criterion that explains when schedule-level fixes are insufficient.

2.2 Positional Encoding for Video-Language Models

Video-language models often apply multi-axis positional encoding across temporal and spatial axes. Prior studies indicate temporal encoding is the main bottleneck for long video. We provide a theoretical reason: temporal displacement grows with clip length, making temporal phase coverage increasingly sparse and error-prone under finite support. Boundary of our contribution: we retain standard multi-axis mRoPE and only intervene on the temporal slice during inference.

2.3 Attention Anisotropy and Spectral Perspectives

Anisotropy in representations and attention has been widely studied via covariance spectra, condition numbers, and effective-rank metrics. Our work extends this line by connecting mRoPE phase dynamics to block-structured covariance coupling, and then connecting spectral peakedness to attention coverage compression. Boundary of our contribution: beyond diagnosis, we derive a direct intervention rule from covariance dynamics.

3 Preliminaries and Problem Setup

3.1 Attention Setup and Notation

For layer l, head h, query index u, and key index v, attention is

$$s_{u,v}^{(l,h)} = \frac{\mathbf{q}_u^\top \mathbf{k}_v}{\sqrt{d_h}}, \qquad \mathbf{a}_{u,:}^{(l,h)} = \mathrm{softmax}(\mathbf{s}_{u,:}^{(l,h)}). \tag{2}$$

We focus on temporal mRoPE channels. Let $\mathbf{z}_{l,h,p} \in \mathbb{R}^{d_t}$ be the temporal feature at position p, where $d_t = 2m$ and each pair of dimensions forms one rotary block. Define centered covariance:

$$\mathbf{M}_{l,h} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} (\mathbf{z}_{l,h,p} - \bar{\mathbf{z}}_{l,h}) (\mathbf{z}_{l,h,p} - \bar{\mathbf{z}}_{l,h})^\top. \tag{3}$$

3.2 Coverage and Spectral Metrics

We track two behavior metrics and three spectral metrics.

Coverage metrics. For attention row $\mathbf{a}_{u,:}$,

$$\mathrm{Span@}p(u) = \min \left\{ |I| : \sum_{v \in I} a_{u,v} \geq p \right\}, \tag{4}$$

and top-k entropy

$$\mathcal{H}_k(u) = - \sum_{v \in \mathrm{TopK}(u)} \hat{a}_{u,v} \log(\hat{a}_{u,v} + \epsilon), \tag{5}$$

where $\hat{a}_{u,v}$ is renormalized over top-k keys. Small Span@p and low $\mathcal{H}_k$ indicate coverage collapse.

Spectral metrics. For covariance $\mathbf{M}$,

$$\Delta_{\mathrm{iso}}(\mathbf{M}) = \left\| \mathbf{M} - \bar{\lambda} \mathbf{I} \right\|_F, \quad \bar{\lambda} = \frac{1}{2m} \mathrm{tr}(\mathbf{M}), \tag{6}$$

$$\kappa(\mathbf{M}) = \frac{\lambda_{\max}(\mathbf{M})}{\lambda_{\min}(\mathbf{M}) + \epsilon}, \tag{7}$$

$$r_{\mathrm{eff}}(\mathbf{M}) = \exp\left(- \sum_i \tilde{\lambda}_i \log(\tilde{\lambda}_i + \epsilon) \right), \quad \tilde{\lambda}_i = \frac{\lambda_i}{\sum_j \lambda_j}. \tag{8}$$

Lower Δ_{iso}, lower κ, and higher r_{eff} indicate flatter spectra.

3.3 mRoPE Temporal Structure

Each temporal rotary block uses frequency ω_i. For two blocks (i, j), relative phase increment is $\Delta_{ij} = \omega_i - \omega_j$. The finite-support cancellation factor is

$$\mathcal{S}_{ij}(N) = \frac{1}{N} \sum_{p=0}^{N-1} e^{\mathrm{i} \Delta_{ij} p}. \tag{9}$$

Its magnitude controls how strongly cross-block interactions survive after aggregation.

4 Theoretical Analysis

4.1 Global Isotropy as Block-Structured Second-Order Decoupling

Assumption Set A (minimal). Temporal channels are organized as m rotary 2×2 blocks; $\mathbf{M}$ in Eq. (3) is finite and positive semidefinite; global isotropy is measured by Eq. (6).

Partition $\mathbf{M}$ into $m \times m$ blocks with 2×2 entries $\mathbf{M}^{(i,j)}$. Then

Theorem 1 (Exact isotropy decomposition).

$$\Delta_{\text{iso}}(\mathbf{M})^2 = \sum_{i=1}^{m} \left\| \mathbf{M}^{(i,i)} - \bar{\lambda}\mathbf{I}_2 \right\|_F^2 + \sum_{i \neq j} \left\| \mathbf{M}^{(i,j)} \right\|_F^2 . \tag{10}$$

Condition. Under Assumption Set A, Eq. (10) is an exact identity. **Result.** Global isotropy decomposes into intra-block isotropy and inter-block decoupling. **Explanation.** Reducing only diagonal anisotropy is insufficient if off-diagonal block energy persists. **Method implication.** Intervention must be selective across heads and target collapse patterns linked to cross-block coupling. *Implication for SPECTRA:* use head-wise degradation sensing rather than uniform perturbation.

4.2 Discrete Phase Cancellation and Cross-Subspace Coupling

Using the geometric-series form,

$$|\mathcal{S}_{ij}(N)| = \frac{1}{N} \left| \frac{\sin(N\Delta_{ij}/2)}{\sin(\Delta_{ij}/2)} \right| . \tag{11}$$

Assume pre-rotation cross-covariance decomposition $\mathbf{B}_p^{(i,j)} = \bar{\mathbf{B}}^{(i,j)} + \Delta\mathbf{B}_p^{(i,j)}$. Then

$$\left\| \mathbf{M}^{(i,j)} \right\|_F \leq \left\| \bar{\mathbf{B}}^{(i,j)} \right\|_F |\mathcal{S}_{ij}(N)| + \frac{1}{N} \sum_{p=0}^{N-1} \left\| \Delta\mathbf{B}_p^{(i,j)} \right\|_F . \tag{12}$$

This bound makes three points explicit: near-frequency pairs cancel slowly, finite N limits cancellation, and non-stationary content leaves residual terms. **When the bound is tight/loose.** The bound is tighter when cross-covariance drift $\Delta\mathbf{B}_p^{(i,j)}$ is small and looser under strong non-stationarity. Hence residual coupling is expected in realistic long-video settings, not an edge case. *Implication for SPECTRA:* intervention strength should increase on heads exhibiting stronger finite-support degeneration.

4.3 Spectral Peakedness Implies Attention Coverage Compression

Under a second-order approximation,

$$\mathrm{Var}(s_{u,v}) = \frac{1}{d_h}\,\mathrm{tr}(\mathbf{\Sigma}_Q\mathbf{\Sigma}_K) \le \frac{1}{d_h}\lambda_{\max}(\mathbf{\Sigma}_Q)\,\mathrm{tr}(\mathbf{\Sigma}_K). \tag{13}$$

When spectra are peaked, $\lambda_{\max}$ dominates and raises logit variance. Higher logit variance increases softmax sharpness, which concentrates probability mass on fewer keys. Consequently Span@p decreases and top-k entropy drops. Therefore spectral flattening is not only correlational; it is a direct mechanism to restore coverage. *Implication for SPECTRA:* monitor spectral collapse with r_{eff} and flatten dominant modes at inference time.

4.4 Why Temporal-Only Intervention in mRoPE

For axis $a \in \{t, h, w\}$, phase excursion is $\Phi_{i,a} = \omega_i \Delta p_a$. In long-video inference, temporal displacement Δp_t grows with clip length, while spatial displacements are bounded by frame size. As a result, temporal channels dominate phase-mismatch risk and residual coupling. This motivates a temporal-only intervention: it offers the highest leverage per perturbation while minimizing side effects on spatial semantics. *Implication for SPECTRA:* apply updates only on temporal dimensions and preserve spatial channels.

5 Method

5.1 Overview

We propose **SPECTRA**, a training-free prefill-time module. Given Q/K states at layer l, SPECTRA performs four steps:

1. Estimate per-head spectral degradation on temporal channels.
2. Compute adaptive gates over layer and head dimensions.
3. Interpolate temporal Q/K with Gaussian anchors using gated strength.
4. Write back only to valid video tokens and temporal dimensions.

The design is strictly plug-and-play: no weight update, no architecture change.
Theory-to-design mapping.

- From Eq. (12): degradation is head-dependent and finite-support dependent, so we use per-head diagnostics instead of uniform perturbation.
- From Eq. (13): dominant eigenmodes drive coverage compression, so we monitor effective rank as a direct collapse signal.
- From Eq. (21): isotropic interpolation contracts dominant directions and lifts weak directions, yielding controlled spectral flattening.
- From temporal phase dominance (Sec. 4.4): intervention is temporal-only to maximize gain and limit semantic side effects.

Implementation Contract. Input: per-layer Q/K tensors, video-token mask, temporal-dimension mask. **Output:** Q/K tensors with only temporal slices of valid video tokens modified. **Execution scope:** prefill only, no parameter update. **Complexity convention:** report added FLOPs as $\mathcal{O}(H\, N_v\, d_t\, r)$ per intervened layer, excluding base attention cost.

5.2 Spectral-Rank-Aware Degradation Signal

For layer l, head h, collect temporal features $\mathbf{X}_{l,h} \in \mathbb{R}^{N_v \times d_t}$. Let top-$r$ singular values be $\sigma_{l,h,1} \geq \cdots \geq \sigma_{l,h,r}$. Define

$$\lambda_{l,h,i} = \frac{\sigma_{l,h,i}^2}{N_v + \epsilon}, \qquad p_{l,h,i} = \frac{\lambda_{l,h,i}}{\sum_{j=1}^{r} \lambda_{l,h,j}}. \tag{14}$$

Head degradation is measured by effective rank:

$$r_{\text{eff}}(l,h) = \exp\left(-\sum_{i=1}^{r} p_{l,h,i} \log(p_{l,h,i} + \epsilon) \right). \tag{15}$$

A smaller $r_{\text{eff}}(l,h)$ means stronger concentration and stronger need for correction. Failure boundary: if singular values are nearly flat but task failure comes from semantics instead of anisotropy, r_{eff} may under-trigger correction.

5.3 Dual Gating: Layer × Head

Layer gate:

$$G_l = \text{clip}\left(1 - \frac{\min_h r_{\text{eff}}(l,h)}{\text{mean}_h r_{\text{eff}}(l,h) + \epsilon}, 0, 1 \right). \tag{16}$$

Head gate:

$$G_{l,h} = \sqrt{\text{clip}\left(\frac{\text{median}_h r_{\text{eff}}(l,h) - r_{\text{eff}}(l,h)}{\text{median}_h r_{\text{eff}}(l,h) - \min_h r_{\text{eff}}(l,h) + \epsilon}, 0, 1 \right)}. \tag{17}$$

Final strength:

$$\alpha_{l,h} = G_l\, G_{l,h}. \tag{18}$$

Monotonicity: with fixed layer statistics, $G_{l,h}$ is non-increasing in $r_{\text{eff}}(l,h)$, so more collapsed heads receive stronger updates. Extreme cases: if all heads are healthy, both gates approach 0; if one head is strongly degraded, that head receives maximal relative strength.

5.4 Training-Free Injection and Complexity

For each temporal vector $\mathbf{x}_{l,h,p}$ (Q or K), sample $\boldsymbol{\eta}_{l,h,p} \sim \mathcal{N}(\mathbf{0}, \sigma_{l,h}^2 \mathbf{I})$ and apply

$$\mathbf{x}'_{l,h,p} = \mathbf{x}_{l,h,p} + \alpha_{l,h}(\boldsymbol{\eta}_{l,h,p} - \mathbf{x}_{l,h,p}) = (1 - \alpha_{l,h})\mathbf{x}_{l,h,p} + \alpha_{l,h}\boldsymbol{\eta}_{l,h,p}. \tag{19}$$

Writeback mask:

$$\mathbf{X}^{\text{out}} = \mathbf{X} + \mathbf{M}_{\text{vid}} \odot \mathbf{M}_{\text{tmp}} \odot (\mathbf{X}' - \mathbf{X}), \tag{20}$$

where $\mathbf{M}_{\text{vid}}$ selects valid video tokens and $\mathbf{M}_{\text{tmp}}$ selects temporal channels.

Proposition 1 (Covariance dynamics under SPECTRA). *Assume $\boldsymbol{\eta}$ is independent isotropic Gaussian. Then*

$$\boldsymbol{\Sigma}'_{l,h} = (1 - \alpha_{l,h})^2 \boldsymbol{\Sigma}_{l,h} + \alpha_{l,h}^2 \sigma_{l,h}^2 \mathbf{I}. \tag{21}$$

Eq. (21) contracts dominant modes and lifts weak modes, which flattens the spectrum and improves coverage robustness. With truncated rank r, per-layer overhead is

$$\mathcal{O}(H\, N_v\, d_t\, r) \tag{22}$$

plus linear-time interpolation. In practice, overhead is modest because SPECTRA is temporal-only, prefill-only, and uses truncated rank.

6 Experiments

6.1 Experimental Setup

This subsection is reserved for datasets, evaluation protocols, baselines, and implementation details.

6.2 Main Results

This subsection is reserved for primary benchmark comparisons under long-video settings.

6.3 Mechanism Validation: Spectrum and Coverage

This subsection is reserved for validating the theoretical chain from spectral flattening to coverage improvement.

6.4 Layer/Head Heterogeneity and Gate Behavior

This subsection is reserved for analyzing gate allocation patterns across layers and heads.

6.5 Ablation Studies

This subsection is reserved for component-wise ablations of degradation signal, dual gating, and injection design.

6.6 Discussion

This subsection is reserved for broader interpretation, practical implications, and failure cases observed in experiments.

7 Limitations

Our analysis is second-order and does not explicitly model higher-order token interactions; this can be tested by adding higher-order diagnostics and checking whether they explain residual errors beyond spectral metrics. The current derivation assumes isotropic Gaussian anchors; this can be relaxed by comparing anisotropic or learned anchors under the same inference budget. Temporal-only intervention is motivated for long-video regimes, but extreme high-motion cases may require coupled temporal-spatial correction; this can be verified by controlled motion-stratified evaluation.

8 Conclusion

We presented a mechanism-first account of long-video attention collapse in mRoPE-based VLMs. The analysis shows that finite-support phase effects induce cross-block coupling, which sharpens spectra and compresses coverage. Based on this chain, SPECTRA provides a training-free, architecture-preserving temporal intervention with explicit covariance-level behavior. Necessity boundary: without controlling cross-block-induced spectral concentration, long-range coverage degradation is expected to persist. Sufficiency boundary: SPECTRA addresses this mechanism directly, but full task optimality may still depend on model semantics and data quality.