BENCHMARKING, DIAGNOSING AND MITIGATING POSITIONAL BIAS IN VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are sensitive to where relevant evidence appears in a video. Existing evaluations often manipulate evidence position by concatenating unrelated clips, which may introduce clip-matching shortcuts and disrupt the contextual relationships needed for reasoning. We investigate temporal positional bias through a unified study of benchmarking, diagnosis, and mitigation. We introduce TempPosBench, which shifts observation windows over continuous videos to vary evidence position while preserving temporal continuity. It comprises 1,322 question–answer instances across 13 tasks spanning perception, reasoning, and streaming awareness. Evaluating seven MLLMs reveals model- and task-specific positional preferences, with temporal reasoning exhibiting the strongest sensitivity. We further examine attention concentrated toward early or late video frames. Targeted prefix calibration and visual-key rotary position embedding (RoPE) compression provide evidence that attention sinks and positional-encoding-induced recency preferences can contribute to positional bias. For mitigation, we extend the benchmark construction to produce positionally balanced training candidates, then select positions that emphasize each model's weaknesses. Fine-tuning on continuous video contexts improves both accuracy and positional robustness on spatial and temporal reasoning compared with concatenated clips. Under comparable training budgets, model-specific selection can also outperform training on all available positions. Fine-tuning additionally yields flatter frame-attention distributions, connecting the training-based mitigation to our attention-level diagnosis. Together, these findings highlight temporal continuity and model-specific positional supervision as complementary ingredients for mitigating temporal positional bias in video understanding.

1 INTRODUCTION

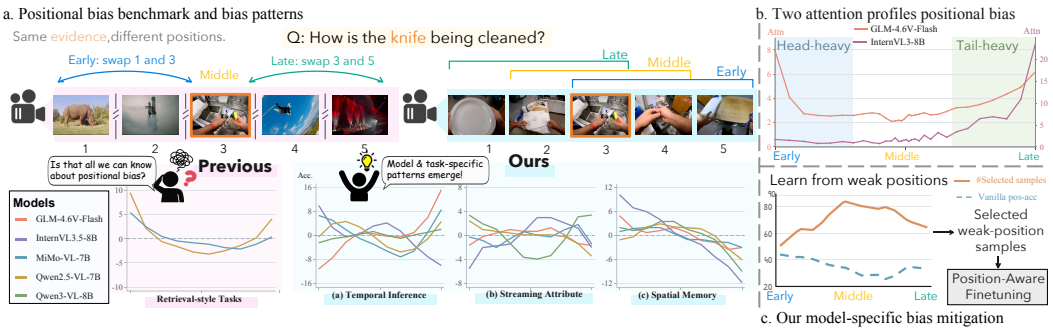

Figure 1: **Overview of our study.** **(a)** Our continuity-preserving construction enables comprehensive evaluation of temporal positional bias in video reasoning. **(b)** We identify two dominant positional attention profiles and probe their effects on temporal positional bias. **(c)** Continuity-preserving, model-specific training data mitigate temporal positional bias.

Multimodal Large Language Models (MLLMs) have shown strong capabilities across video perception, reasoning, and streaming understanding. Yet their performance can vary substantially depending on where relevant evidence appears in a video, raising concerns about temporal positional

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

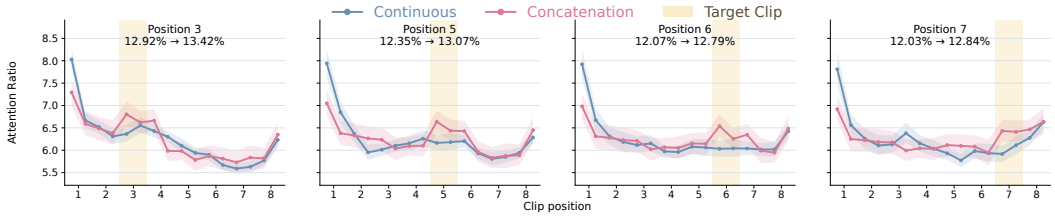

Figure 2: Attention under continuous and concatenated video contexts at four selected evidence positions (3, 5, 6, and 7) where target-aligned elevations are more pronounced under concatenation. Yellow bands mark the target clip. Results for all eight positions are provided in Figure 9 in the appendix.

bias. Prior studies in text LLMs often examine this issue in retrieval-style settings, placing relevant evidence among distractors. Constructing comparable position-controlled inputs is more challenging for video: moving an evidence clip can disrupt the temporal relationships with its surrounding context. Existing video studies therefore adopt concatenation-based constructions, inserting a target clip among unrelated background clips to vary its position.

However, this construction may introduce a **semantic shortcut**. Figure 2 compares attention under concatenated and continuous video contexts. Under concatenation, attention exhibits more pronounced local elevations around the evidence clip and tracks its position more closely. These elevations are less pronounced under continuous context, suggesting that concatenation may make the evidence clip easier to isolate from its surroundings. Positional evaluation may therefore become entangled with clip localization enabled by the construction itself.

This observation raises a central question: *Does preserving temporal continuity enable more faithful benchmarking, diagnosis, and mitigation of temporal positional bias?* Specifically, we ask whether continuous video contexts better preserve the intended reasoning demands during positional evaluation, enable attention-level diagnosis without concatenation-induced shortcuts, and provide more effective training data for positional-bias mitigation.

We therefore introduce **TempPosBench**, a position-controlled yet continuity-preserving benchmark that varies the temporal position of the target clip while retaining the original temporal continuity. TempPosBench spans six capability categories across perception, reasoning, and streaming awareness, whose intended task semantics remain intact under this design.

Moving from performance-level bias to internal model behavior, we further examine question-to-frame attention allocation. Across models, we observe **sharp positional attention profiles**(Fig 1(b)), characterized by distinct **head-heavy** and/or **tail-heavy** concentration. These signatures motivate two corresponding hypotheses: early-token attention sinks for the head-heavy pattern, and a positional-encoding-induced preference for later tokens for the tail-heavy pattern. We probe both through targeted interventions.

Moving beyond pattern-specific corrections, we ask whether a **positionally balanced training-data design can reduce positional performance range**. We extend the same continuity-preserving principle from evaluation to training by exposing the model to target clips across different temporal positions. Compared with concatenation-based training constructed from the same VQA data, continuity-preserving training achieves both higher task performance and a smaller positional range on held-out TempPosBench. These results show that positional coverage alone is insufficient: preserving temporal continuity provides more effective supervision for positional bias mitigation. While continuity-preserving training broadly improves performance over the vanilla model, its advantage over concatenation-based training is most pronounced in spatial and temporal reasoning, suggesting that preserving temporal continuity becomes particularly important for reasoning-intensive tasks.

Our contributions are threefold:

- **Broad positional-bias benchmarking for video understanding.** We introduce **TempPosBench**, which, to the best of our knowledge, is the first positional-bias benchmark to jointly evaluate perception, reasoning, and streaming awareness. This breadth is enabled by a continuity-preserving design that controls target-clip position without altering the original temporal structure, preserving the intended semantics of diverse task types.

- **Attention-level diagnosis of positional bias.** We find that attention remains sharply biased toward the beginning or end of the video instead of following the target clip across positions, forming distinct **head-heavy** and **tail-heavy** patterns across models. Interventions targeting early-token attention sinks and the positional-encoding-induced preference for later tokens attenuate the head-heavy and tail-heavy patterns, respectively, establishing a link between these attention profiles and positional bias.
- **Positionally balanced training data for bias mitigation.** We show that effective mitigation of temporal positional bias in video reasoning requires both model-specific supervision and temporal continuity in the training videos. In contrast, the semantic shortcut introduced by concatenation-based constructions makes positional-bias mitigation substantially harder on reasoning-intensive tasks.

2 Related Work

2.1 Positional Bias in Language and Multimodal Models

Long-context benchmarks reveal difficulties in retrieving and reasoning over long text (Bai et al., 2024; Hsieh et al., 2024a; Kuratov et al., 2024; Bai et al., 2025c) and multimodal inputs (Song et al., 2024; Wu et al., 2024; Zhou et al., 2025; Fu et al., 2025). Positional-bias studies show that LLM performance varies with evidence location (Liu et al., 2024; Song et al., 2025), while studies of multi-hop question answering show that models are also sensitive to the spacing and order of supporting evidence (Baker et al., 2024; Yu et al., 2025a). Researchers also report positional effects in multi-image understanding, multimodal retrieval, and multi-video summarization (Tan et al., 2024; Tian et al., 2025; Yao et al., 2025; Liu et al., 2026a; Xu et al., 2026). For video, VideoNIAH uses synthetic retrieval, ordering, and counting tasks (Zhao et al., 2024), while Video-LevelGauge tests positional bias across multiple tasks by inserting standardized probe clips into customized contexts (Xia et al., 2026). However, concatenating semantically unrelated clips, as in Video-LevelGauge (Xia et al., 2026),disrupts temporal continuity and may reduce distinct tasks to clip retrieval, limiting task-specific positional-bias analysis.

2.2 Diagnosing and Mitigating Positional Bias

Studies link positional bias to causal attention, positional encodings, and their interaction across layers (Wang et al., 2025b; Wu et al., 2025; Barbero et al., 2024), with additional effects from LayerNorm and position-dependent hidden states (Kim et al., 2026; Yu et al., 2025b). One line of work examines attention sinks, where tokens attract high attention despite limited semantic relevance (Xiao et al., 2024; Gu et al., 2025; Ran-Milo et al., 2026). Multimodal studies also identify visual attention sinks and use attention redistribution or layer-wise gating to improve visual information use (Kang et al., 2025; Choi et al., 2026). Another line targets positional representations through hidden-state scaling and positional-encoding calibration (Yu et al., 2025b; Zhang et al., 2024b; Zehle & Aßenmacher, 2026; Wang et al., 2026). For video, VideoRoPE and VRoPE revise spatiotemporal position layouts and frequency allocation (Wei et al., 2025; Liu et al., 2025), while PAS reduces sensitivity to small temporal shifts through phase aggregation (Sun et al., 2026). Researchers also calibrate positional attention (Hsieh et al., 2024b), modify attention across documents or images (Wang et al., 2025b; Tian et al., 2025), and rebalance frame-level attention to mitigate video hallucinations (Liu et al., 2026b). However, attention patterns alone do not establish their effects on predictions (Liu et al., 2022; Xian et al., 2026). We probe attention sinks and positional-encoding effects in continuous videos and investigate positionally balanced training that preserves temporal continuity.

3 Approach: Benchmarking, Diagnosis, and Mitigation

3.1 Benchmark Design

We design **TempPosBench** to evaluate how model performance changes when the same question-answering evidence appears at different positions within a continuous video. It combines context-

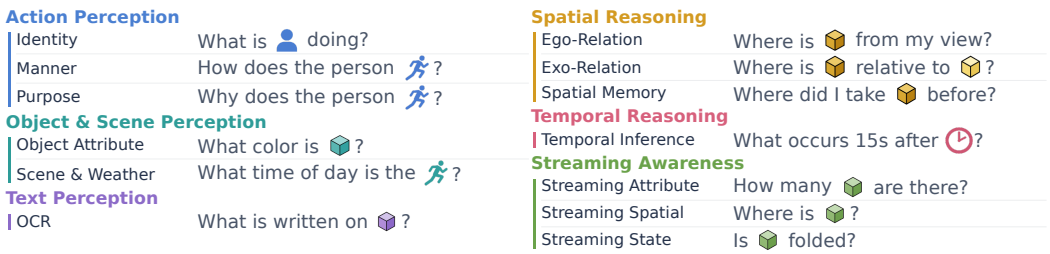

Figure 3: TempPosBench evaluates temporal positional bias across 13 tasks spanning six capabilities.

window shifting with tasks covering perception, reasoning, and streaming awareness, providing a shared evaluation basis for diagnosing and mitigating temporal positional bias.

Context-Window Shifting.

For each VideoQA instance, let V^{src} denote the source video and $V^{\text{tgt}} = V^{\text{src}}[t_s : t_e]$ the target interval containing the evidence needed to answer question q. We extract fixed-duration windows with different start times, placing the complete target interval at different positions within each input. Contiguous extraction preserves the original temporal order and spacing without introducing artificial concatenation boundaries.

For window duration L, target duration $d = t_e - t_s < L$, and normalized placement $p_i \in [0, 1]$, the window boundaries are

$$l_i = t_s - p_i(L - d), \qquad r_i = l_i + L,$$

yielding $V_i = V^{\text{src}}[l_i : r_i]$. Placements $p_i = 0$ and $p_i = 1$ place the target interval at the beginning and end of the input, respectively. Each QA yields K variants at predefined placements $\mathcal{P} = \{p_1, \ldots, p_K\}$. All windows must remain within the source video, and every placement must be evaluated on the same eligible QA instances. Across variants, only target placement and surrounding context change; the target interval and window duration remain fixed.

We update window-dependent temporal references in questions and answers while leaving offsets between fixed events unchanged. These edits preserve the referenced events, question semantics, and correct-answer identity.

Task Design.

We organize thirteen tasks into six task families to examine positional sensitivity across visual content and response settings. Figure 3 presents the taxonomy and example questions.

Eligible instances must provide a recoverable source video and a temporally localized interval containing all necessary evidence. Questions must remain well-defined across windows, referring to the same event, object, or target time with a valid answer. For multi-moment tasks such as spatial memory, the target interval must cover all required observations. We exclude instances whose intended task changes under window shifting, such as forecasting questions for which added future observations reveal the answer.

For streaming awareness, the model should emit `<attention>` and answer once sufficient evidence has appeared in the observed video, and emit `<silence>` otherwise.

Evaluation metrics. For each model m and task t, let $A_{m,t}(p)$ denote accuracy at placement $p \in \mathcal{P}$. We report mean accuracy and positional range:

$$\bar{A}_{m,t} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} A_{m,t}(p), \qquad R_{m,t} = \max_{p \in \mathcal{P}} A_{m,t}(p) - \min_{p \in \mathcal{P}} A_{m,t}(p).$$

We report both because a small range may also reflect uniformly poor performance across positions.

3.2 Diagnosis

We focus our diagnosis on **Temporal Reasoning**, as it is a fundamental capability in video understanding and exhibits the largest positional bias in our benchmark. Building on the context-window-shifting protocol in Section 3.1, we examine how frame-level attention changes with evidence position and probe two potential mechanisms of temporal positional bias: early-token attention sinks via

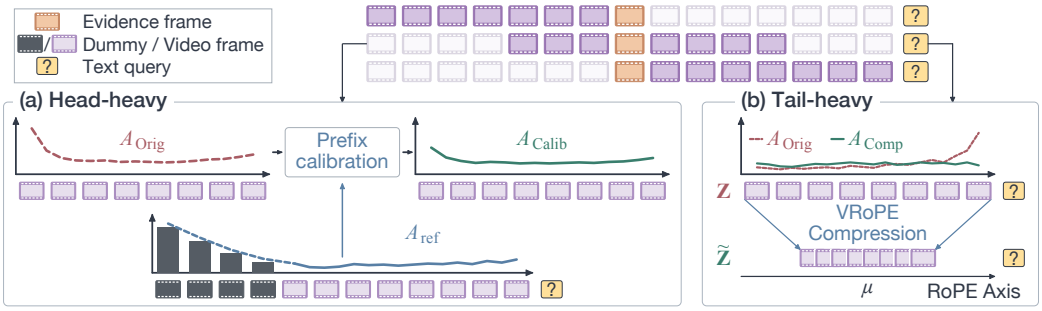

Figure 4: Diagnostic interventions for two positional attention patterns. (a) Prefix calibration derives a reference attention profile A_{ref} from a black-frame-prefixed pass. (b) Visual-key RoPE compression contracts native coordinates $\mathbf{z}$ toward their mean $\boldsymbol{\mu}$. Curves are schematic.

buffer-frame calibration, and positional-encoding effects via visual-key RoPE compression (Figure 4). We measure text-to-frame attention during input processing, averaged over all heads and post-video prompt tokens, and align the resulting profiles by within-window frame position across target placements. Behavioral effects are evaluated using mean task performance Am, t and positional range Rm, t defined in Section 3.1.

Diagnosing head-heavy attention. We hypothesize that elevated attention to early video frames partly reflects a structural preference for earlier positions rather than question relevance. To probe this, we prepend question-irrelevant black frames and observe that attention remains concentrated on these newly introduced early frames instead of shifting to the original frames containing relevant evidence.

Our prefix calibration interpolates the frame-attention distribution A_{orig} on the original frame sequence with this reference:

$$A_{\text{cal}} = (1 - \lambda)A_{\text{orig}} + \lambda A_{\text{ref}},$$

where $\lambda \in [0, 1]$ controls the calibration strength. We apply calibration to the layer group whose attention profile changes most after prepending black frames. This intervention tests whether reducing the hypothesized structural positional bias is accompanied by a smaller temporal positional range.

Diagnosing tail-heavy attention. We test whether rotary position embedding (RoPE) contributes to the preference for later visual tokens in text-to-visual attention. RoPE exhibits long-term decay (Su et al., 2024), which may favor later visual tokens that are positionally closer to post-video text queries. We therefore compress visual-key RoPE coordinates toward their mean for post-video text queries, while retaining native RoPE for all other attention interactions to preserve inter-frame processing. Let $\mathcal{V}$ and $\mathcal{T}_{\text{post}}$ denote visual and post-video text tokens, respectively, and let $\mathbf{z}_j$ be the native RoPE coordinates of token j. We define

$$\widetilde{\mathbf{z}}_j = \boldsymbol{\mu} + \alpha(\mathbf{z}_j - \boldsymbol{\mu}), \qquad \boldsymbol{\mu} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \mathbf{z}_j,$$

where $\alpha \in [0, 1]$ controls the retained coordinate spread and applies to the model's native one-dimensional or spatiotemporal coordinates.

For attention from query i to key j,

$$\mathbf{k}_j^{(i)} = \begin{cases} R(\widetilde{\mathbf{z}}_j)\mathbf{k}_j, & i \in \mathcal{T}_{\text{post}},\ j \in \mathcal{V}, \\ R(\mathbf{z}_j)\mathbf{k}_j, & \text{otherwise}, \end{cases}$$

where $\mathbf{k}_j$ is the pre-RoPE key and R is the native rotary transform. Queries, values, and masks remain unchanged.

3.3 Mitigation

We hypothesize that temporal positional bias may partly reflect an imbalanced distribution of QA-related evidence across temporal positions in training data. Our mitigation has two stages: continuity-preserving candidate construction with balanced positional coverage, followed by model-specific selection that prioritizes weak positions of the target model.

Positionally balanced candidate construction. Following the context-window-shifting protocol in Section 3.1, we construct one example at each prescribed placement for every training QA, giving all placements equal representation in the candidate pool. In contrast, positional coverage through concatenating unrelated clips may encourage clip-matching shortcuts rather than reasoning over continuous temporal context. As a control for dataset familiarity, we also fine-tune on the corresponding raw videos without position-balanced window construction, testing whether the gains arise merely from exposure to the dataset semantics.

Model-specific positional selection. Because each model exhibits its own positional-bias profile, we tailor the training positions to each target model by prioritizing placements where it performs poorly. For model m and QA instance q, we score each placement p using the answer-probability margin:

$$s_{m,q}(p) = P_m(a_q^\star \mid V_{q,p}, q) - \max_{a \neq a_q^\star} P_m(a \mid V_{q,p}, q),$$

where $V_{q,p}$ denotes the video window and $a_q^\star$ the correct answer. We retain the k lowest-scoring placements:

$$\mathcal{S}_{m,q} = \mathrm{BottomK}_{p \in \mathcal{P}}\big(s_{m,q}(p), k\big).$$

This preserves every QA instance while adapting its temporal placements to the target model.

4 Experiments

4.1 Experimental Setup

Data. Our benchmark comprises 1,322 QA instances from 726 source videos across CG-Bench (Chen et al., 2025), HD-EPIC (Perrett et al., 2025), EMCompress (Fan et al., 2026), and IPIBench (Li et al., 2026). Evidence timestamps are directly inherited from the source benchmarks. For fine-tuning experiments, we use a 6:4 train-test split at the QA level, keeping all variants of the same QA in the same split.

Models and inference. We evaluate seven models: GLM-4.6V-Flash (Hong et al., 2025), InternVL3-8B (Zhu et al., 2025), InternVL3.5-8B (Wang et al., 2025a), MiMo-VL-7B (Yue et al., 2025), LLaVA-Video-7B (Zhang et al., 2024a), Qwen2.5-VL-7B (Bai et al., 2025b), and Qwen3-VL-8B (Bai et al., 2025a). We use each model's native video-input pipeline with 40 frames over 80-100 s observation windows and greedy decoding without thinking mode. Prompts for standard and streaming evaluation are provided in the supplementary materials. For InternVL3-8B and GLM-4.6V-Flash in diagnosis, we bypass GLM's native video processor and instead represent each video as a sequence of consecutive image frames, enabling direct control over RoPE indices and attention computation. Prefix calibration is applied to layers 7–9 and 18–23 with $\lambda = 0.50$ and 0.75, while V-RoPE compression uses $\alpha = 0.25$ and 0.50 (see Appendix B.1 for selection details).

Mitigation training. We perform supervised fine-tuning with cross-entropy loss using MS-Swift and LoRA (rank 16, scaling factor 32), while freezing the visual encoder and multimodal projector. We use AdamW with a learning rate of 5×10^{-5} and an effective batch size of 8. Training epochs are adjusted across position subsets to approximately match the number of optimizer updates; full optimization details are provided in the supplementary materials.

4.2 Benchmarking: Temporal Positional Bias in Video Reasoning

Table 1 reports capability-level mean accuracy and positional ranges on TempPosBench. Figure 1(a) shows representative position-wise profiles across three non-retrieval tasks, expressed as deviations from each model's across-position mean.

Temporal reasoning exhibits the strongest positional sensitivity. Position-dependent performance variation appears across all evaluated models and task families. Temporal reasoning exhibits the largest positional range for every model (9.68–30.32 percentage points), highlighting its particular sensitivity to where evidence appears within the observed sequence.

Higher accuracy does not necessarily imply greater positional robustness. In Streaming Awareness, InternVL3.5 and Qwen3-VL achieve similar mean joint accuracies (44.51% and 44.14%), but their positional ranges differ (9.90 versus 7.14 percentage points; Table 1). Conversely, a relatively

Table 1: Main results on TempPosBench. Bold marks the best value per column.

Model	Action Perception		Obj./Scene Perception		Text Perception		Spatial Reasoning		Temporal Reasoning		Streaming Awareness	
	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range	Mean	Range
Qwen3-VL-8B	39.45	8.78	58.39	8.10	60.82	5.48	**41.20**	8.89	**61.62**	**9.68**	44.14	7.14
GLM-4.6V-Flash	36.59	8.59	55.45	7.40	56.30	9.59	37.04	9.11	49.18	29.68	31.98	**4.79**
LLaVA-Video-7B	33.78	6.56	46.05	6.27	50.96	**4.11**	30.71	10.44	39.57	30.32	31.88	6.20
InternVL3-8B	**41.46**	**5.81**	**58.73**	7.49	**63.15**	**4.11**	34.73	7.56	36.59	18.71	35.76	9.97
InternVL3.5-8B	40.37	10.59	55.64	7.14	58.08	12.33	39.07	13.22	37.11	22.58	**44.51**	9.90
MiMo-VL-7B	36.66	7.00	50.77	**4.79**	47.40	5.48	32.42	7.78	37.03	23.87	43.19	8.12
Qwen2.5-VL-7B	40.47	7.70	53.91	6.53	53.42	8.22	37.11	**7.22**	41.28	18.71	18.88	5.83

small range can also accompany poor accuracy. Mean accuracy and positional range must therefore be considered together when assessing reliable performance across positions.

Temporal positional bias is model- and task-specific. Figure 1(a) shows contrasting preferences on Temporal Inference: GLM favors later evidence, whereas InternVL3.5 favors earlier evidence; approximately U-shaped profiles also appear. These preferences change across tasks: on Spatial Memory, all plotted models perform below their respective means at the final position, including those favoring later evidence on Temporal Inference. Attribute (Streaming Awareness) further exhibits contrasting U-shaped and inverted-U-shaped profiles across models. This variation motivates a case-specific attention diagnosis that accounts for both the model and the task.

4.3 Diagnosis: Position-Dependent Attention Distributions

Figure 1(b) shows distinct frame-attention profiles: GLM-4.6V-Flash exhibits elevated attention at both ends of the video, whereas InternVL3-8B exhibits predominantly tail-heavy attention without a comparable early-frame peak. We use prefix calibration and visual-key RoPE compression as pattern-specific interventions to probe how these attention profiles relate to positional bias.

Prefix calibration. Table 2 shows that prefix calibration reduces GLM-4.6V-Flash's positional range from 18.19 to 8.37, with a slight decrease in mean accuracy. In contrast, InternVL3-8B shows only a marginal range reduction. This model-dependent response is consistent with a link between early-position attention concentration and positional bias in this setting.

Table 2: **Diagnosis result on Temporal Reasoning.**

Model	Setting	Acc. ↑	Range ↓
InternVL3-8B	Vanilla	34.45	14.40
	Calibrated	34.87	13.68
	Compressed	38.08	12.20
GLM-4.6V-Flash	Vanilla	34.24	18.19
	Calibrated	33.23	8.37
	Compressed	36.96	15.46

Visual-key RoPE compression. To probe tail-heavy attention, we compress visual-key RoPE coordinates to reduce differences in query-relative positional offsets across frames. Figure 8 shows that attention becomes less concentrated toward later frames after compression, supporting our RoPE-recency induced tail-heavy attention hypothesis. Further, this redistribution is accompanied by higher accuracy and smaller positional ranges for both models on Temporal Reasoning (Table 2), supporting a connection between tail-heavy attention and temporal positional bias.

Additional ablations in the appendix (Table 7) support restricting compression to text-to-visual attention and show that the range reduction in GLM's M-RoPE is not attributable to the temporal axis alone. Together, these findings support RoPE-induced recency preference as a potential contributor to tail-heavy attention and temporal positional bias.

4.4 Mitigation: Reducing Temporal Positional Bias

Beyond the model–task-specific diagnostic interventions, we evaluate the training-based mitigation in Section 3.3 across tasks, comparing data constructions and positional selection strategies.

Temporal continuity matters for positional-bias mitigation in reasoning. We compare fine-tuning on continuous windows and concatenated clips. This tests whether the closer alignment between attention and evidence-clip positions observed under concatenation translates into better mitigation. Table 3 shows that training on continuous windows yields higher overall mean accuracy (61.88 versus 54.64) and a smaller positional range (17.06 versus 20.98) than training on concate-

Table 3: Bias mitigation results with raw, concatenated, and continuity-preserving training videos.

Setting	Metric	Overall	Action Perception	Obj./Scene Perception	Text Perception	Spatial Reasoning	Temporal Reasoning
Vanilla	Mean	38.91	34.55	44.86	55.45	29.17	30.51
	Range	31.49	26.59	24.76	37.17	32.18	36.76
Raw video	Mean	20.59 (-18.32)	19.86 (-14.69)	18.58 (-26.28)	19.83 (-35.62)	21.11 (-8.06)	23.55 (-6.96)
	Range	20.06 (-11.43)	20.70 (-5.89)	20.03 (-4.73)	23.90 (-13.27)	22.33 (-9.85)	13.36 (-23.40)
Concat	Mean	54.64 (+15.73)	46.55 (+12.00)	58.59 (+13.73)	73.52 (+18.07)	49.01 (+19.84)	45.54 (+15.03)
	Range	20.98 (-10.51)	18.77 (-7.82)	15.62 (-9.14)	13.72 (-23.45)	20.66 (-11.52)	36.10 (-0.66)
Continuous	Mean	**61.88** (+22.97)	45.63 (+11.08)	57.09 (+12.23)	78.69 (+23.24)	**53.19** (+24.02)	74.79 (+44.28)
	Range	**17.06** (-14.43)	16.90 (-9.69)	14.50 (-10.26)	20.10 (-17.07)	**17.91** (-14.27)	15.88 (-20.88)

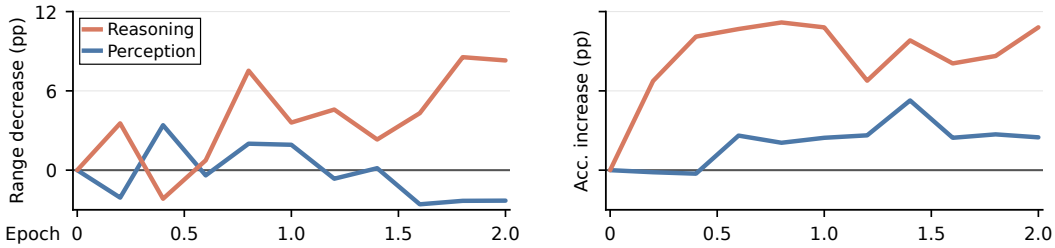

Figure 5: Training dynamics on continuous versus concatenated videos for perception and reasoning. Positional range reduction (left) and accuracy gain (right) are measured in percentage points; positive values favor training on continuous videos.

nated clips. The advantage is particularly pronounced in temporal reasoning, with accuracy of 74.79 versus 45.54 and a range of 15.88 versus 36.10. Spatial reasoning also improves on both metrics. Perception results are more mixed: training on continuous windows does not consistently outperform training on concatenated clips in either accuracy or bias mitigation. Figure 5 traces how the effects of the two data constructions evolve over successive training epochs. As training progresses, training on continuous windows sustains a larger accuracy advantage for reasoning than for perception, while its range-reduction advantage for reasoning becomes more pronounced toward the later epochs. In contrast, perception shows comparatively modest accuracy gains and no sustained advantage in range reduction. These trajectories suggest that preserving continuous context benefits spatial and temporal reasoning, whereas concatenation may offer a clip-matching shortcut that hinders the learning of these reasoning capabilities.

Model-specific positional selection improves bias mitigation For each QA, we compare model-specific and random position selection against using all available positions. Every setting retains the same QA instances, with approximately matched optimizer steps. For the same number of positions per QA, model-specific selection consistently achieves higher mean accuracy than random selection (Table 4). Model-specific selection of four or six positions also yields smaller positional ranges than both random selection and full-position training. With six positions, mean accuracy reaches 57.48 and the range falls to 15.05, outperforming training on all positions (56.41 and 16.94, respectively). Expanding the selection from six to eight positions worsens the positional range by 2.94 pp, despite a modest 0.58 pp gain in mean accuracy. This suggests that adding easier positions may dilute the training focus on the model's weaknesses. The model-specific selection strategy also improves mean accuracy and reduces positional ranges on InternVL3-8B and GLM-4.6V-Flash relative to their vanilla baselines (Table 5). Under comparable training budgets, emphasizing model-specific weaknesses can be more effective than training on all available positions.

Positional-bias mitigation needs positionally balanced training data. To test whether these gains merely reflect familiarity with the training videos, we additionally fine-tune on the corresponding raw videos(Table 3). Although the positional range decreases, mean accuracy drops sharply from 38.91 to 20.59, making this setting qualitatively incomparable to the two positionally balanced variants, which improve both accuracy and positional robustness. These results suggest that exposure to the same video content alone is insufficient for effective positional-bias mitigation.

Table 4: Training-position selection under comparable training steps on Mimo-VL. The highest mean accuracy and lowest positional ranges are bolded.

Metric	Selection	Reference		Positions per QA			
		Vanilla	Full	2	4	6	8
Mean ↑	Random	36.68	56.41	52.55	54.46	55.37	57.41
	Model-specific			53.82	57.01	57.48	**58.06**
Range ↓	Random	29.98	16.94	16.54	18.41	16.10	17.44
	Model-specific			17.22	16.03	**15.05**	17.99

Table 5: Mitigation across models with six model-specific positions per QA.

Model	Setting	Mean ↑	Range ↓
MiMo-VL-7B	Vanilla	36.68	29.98
	Model-specific	**57.48**	**15.05**
InternVL3-8B	Vanilla	40.75	19.85
	Model-specific	**55.91**	**16.83**
GLM-4.6V-Flash	Vanilla	40.26	20.05
	Model-specific	**53.34**	**13.46**

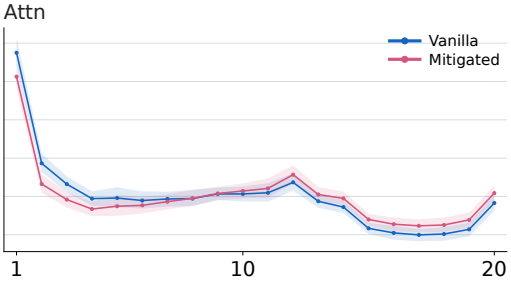

Figure 6: Attention of Mimo-VL-7B vanilla and Mitigated, on Spatial Reasoning.

Attention after fine-tuning. By comparing the model's attention distributions on Spatial Reasoning tasks before and after fine-tuning, we observe a modest reduction in early-position concentration (Figure 12). Although the overall profile remains similar, this partial redistribution qualitatively echoes our attention-level diagnosis in Section 4.3.

5 Conclusion

We investigated temporal positional bias in video understanding through benchmarking, attention-level diagnosis, and training-based mitigation. TempPosBench reveals model- and task-specific positional preferences, with temporal reasoning showing particularly strong sensitivity. Targeted interventions provide evidence that early-position attention sinks and RoPE-related recency preferences can contribute to this bias. Training on continuous video contexts improves spatial and temporal reasoning compared with concatenated clips. We also observe a partial reduction in structural attention bias after fine-tuning, qualitatively echoing our diagnosis. Together, these findings highlight the importance of temporal continuity for evaluating and mitigating positional bias. Within this continuity-preserving setting, broader positional coverage does not necessarily yield better mitigation: under comparable training budgets, selecting positions that emphasize model-specific weaknesses can outperform training on all positions.

AI Use Statement

We used generative AI tools to assist with manuscript drafting, translation, and language refinement, as well as discussions of methodology and result interpretation. These tools also assisted with LaTeX formatting and figure and table preparation. The authors reviewed and revised the AI-assisted material and take responsibility for the final manuscript, including its scientific claims and reported results.

Ethics Statement

This work studies temporal positional bias using existing publicly available video-QA datasets. We do not identify additional ethical concerns specific to the benchmark construction, model evaluation, or mitigation procedures described in this paper.

Reproducibility Statement

Code is included with this submission. Section 3 describes benchmark construction, evaluation metrics, diagnostic interventions, and training-data selection. Section 4.1 reports the evaluated models, inference settings, data splits, and fine-tuning hyperparameters. Appendix B.1 documents diagnostic layer and hyperparameter selection, while Appendix A.1 provides additional task-level results and ablations.

References

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025c.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa. *arXiv preprint arXiv:2412.10079*, 2024.

Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful? *arXiv preprint arXiv:2410.06205*, 2024.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2026.

Zheyu Fan, Jiateng Liu, Yuji Zhang, Zihan Wang, Yi R Fung, Manling Li, and Heng Ji. Emcompress: Video-llms with endomorphic multimodal compression. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 137–162, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations*, volume 2025, pp. 97114–97144, 2025.

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

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024a.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, et al. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14982–14995, 2024b.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Junu Kim, Xiao Liu, Zhenghao Lin, Lei Ji, Yeyun Gong, and Edward Choi. Layernorm induces recency bias in transformer decoders. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28638–28652, 2026.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106519–106554, 2024.

Jinzhao Li, Yinuo Chen, Wenxuan Song, Yijia Lei, Yichi Zhang, Honglei Yan, Panwang Pan, and Miao Liu. Ipibench: Evaluating interactive proactive intelligence of mllms under continuous streams. *arXiv preprint arXiv:2605.27074*, 2026.

Jieyuan Liu, Jianyang Gu, Shijie Chen, Jefferson Chen, and Zhen Wang. Lost at the end: Primacy bias in multimodal retrieval-augmented question answering. *arXiv preprint arXiv:2606.16494*, 2026a.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. Rethinking attention-model explainability through faithfulness violation test. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 13807–13824, 2022.

Zijian Liu, Sihan Cao, Pengcheng Zheng, Kuien Liu, Caiyan Qin, Xiaolin Qin, Jiwei Wei, and Chaoning Zhang. Relaxing anchor-frame dominance for mitigating hallucinations in video large language models. *arXiv preprint arXiv:2604.12582*, 2026b.

Zikang Liu, Longteng Guo, Yepeng Tang, Tongtian Yue, Junxian Cai, Kai Ma, Qingbin Liu, Xi Chen, and Jing Liu. VRoPE: Rotary position embedding for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14460–14472, 2025. doi: 10.18653/v1/2025.emnlp-main.731.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, et al. Hd-epic: A highly-detailed egocentric video dataset. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913. IEEE, 2025.

Yuval Ran-Milo, Hila Ofek, and Shahar Mendel. A mechanistic account of attention sinks in gpt-2: One circuit, broader implications for mitigation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 90–98, 2026.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

Mingyang Song, Mao Zheng, and Xuan Luo. Counting-stars: A multi-evidence, position-aware, and scalable benchmark for evaluating long-context large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 3753–3763, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Bowen Sun, Yujun Cai, Ming-Hsuan Yang, Hang Wu, and Yiwei Wang. Pas: A training-free stabilizer for temporal encoding in video llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14480, 2026.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*, 2024.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10599–10609. IEEE, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Zhenghua Wang, Yiran Ding, Changze Lv, Yixin Wu, Tianlong Li, Zhibo Xu, Muling Wu, Tianyuan Shi, Shizheng Li, Qi Qian, et al. Mitigating position bias in transformers via layer-specific positional embedding scaling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21084–21098, 2026.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models:a mechanistic approach. In *International Conference on Learning Representations*, 2025b.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. *arXiv preprint arXiv:2502.01951*, 2025.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. Videolevelgauge: Investigating contextual positional bias in video language models. In *International Conference on Learning Representations*, volume 2026, pp. 45877–45905, 2026.

Mingtao Xian, Yifeng Yang, Qinying Gu, Xinbing Wang, and Nanyang Ye. Logit-attention divergence: Mitigating position bias in multi-image retrieval via attention-guided calibration. *arXiv preprint arXiv:2605.11591*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Huangchen Xu, Yuan Wu, and Yi Chang. A systematic evaluation of positional bias in multi-video summarization with mllms. *arXiv preprint arXiv:2606.04596*, 2026.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15183–15193, 2025. doi: 10.18653/v1/2025.emnlp-main.768. URL https://aclanthology.org/2025.emnlp-main.768/.

Sangwon Yu, Ik-hwan Kim, Jongyoon Song, Saehyung Lee, Junsung Park, and Sungroh Yoon. Unleashing multi-hop reasoning potential in large language models through repetition of misordered context. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6435–6455, 2025a.

Yijiong Yu, Huiqiang Jiang, Xufang Luo, Qianhui Wu, Chin-Yew Lin, Dongsheng Li, Yuqing Yang, Yongfeng Huang, and Lili Qiu. Mitigate position bias in llms via scaling a single hidden states channel. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6092–6111, 2025b.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Tom Zehle and Matthias Aßenmacher. Can calibration of positional encodings enhance long context utilization? In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2268–2280, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024a.

Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. *Advances in Neural Information Processing Systems*, 37: 60755–60775, 2024b.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. *arXiv preprint arXiv:2406.09367*, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

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

A BENCHMARK

A.1 TASK-LEVEL BENCHMARK RESULTS

Table 6: Task-level benchmark performance on TempPosBench.

Model Mean / Range	Action Perception			Object & Scene Perception		Text Perception
	Identity	Manner	Purpose	Object Attribute	Scene & Weather	OCR
Qwen3-VL-8B	44.44/13.33	27.80/ 6.00	**46.10**/ 7.00	54.35/ 5.22	**62.44**/10.98	60.82/ 5.48
GLM-4.6V-Flash	41.67/ 7.78	32.70/ 9.00	35.40/ 9.00	52.00/ 8.70	58.90/ **6.10**	56.30/ 9.59
LLaVA-Video-7B	24.44/ 6.67	31.10/ 7.00	45.80/ 6.00	40.26/ 5.22	51.83/ 7.32	50.96/ **4.11**
InternVL3-8B	**47.78**/ **4.44**	35.90/ 7.00	40.70/ 6.00	**56.00**/ 5.22	61.46/ 9.76	**63.15**/ **4.11**
InternVL3.5-8B	42.22/ 7.78	33.80/ 8.00	45.10/16.00	50.17/ 6.96	61.10/ 7.32	58.08/12.33
MiMo-VL-7B	39.67/10.00	**38.00**/ **4.00**	32.30/ 7.00	50.43/ **3.48**	51.10/ **6.10**	47.40/ 5.48
Qwen2.5-VL-7B	39.11/11.11	36.60/ 7.00	45.70/ **5.00**	48.43/ 6.96	59.39/ **6.10**	53.42/ 8.22

Model Mean / Range	Spatial Reasoning			Temporal Reasoning	Streaming Awareness		
	Ego-Relation	Exo-Relation	Spatial Memory	Temporal Inference	Attribute	Spatial	State
Qwen3-VL-8B	20.80/ 6.00	**42.80**/ 6.67	**60.00**/14.00	**61.62**/ **9.68**	38.94/10.59	32.34/ 6.54	**61.14**/ **4.29**
GLM-4.6V-Flash	18.00/ 6.00	42.13/ 9.33	51.00/12.00	49.18/29.68	35.76/ **4.71**	33.18/ 4.67	27.00/ 5.00
LLaVA-Video-7B	18.50/ 7.00	30.93/ **5.33**	42.70/19.00	39.57/30.32	25.06/ 7.06	24.86/ 6.54	45.71/ 5.00
InternVL3-8B	25.60/ 8.00	39.20/ 6.67	39.40/ 8.00	36.59/18.71	24.59/ 9.41	25.61/11.21	57.07/ 9.29
InternVL3.5-8B	**30.00**/ 7.00	36.40/10.67	50.80/22.00	37.11/22.58	36.12/10.59	**39.35**/ 8.41	58.07/10.71
MiMo-VL-7B	26.20/ 8.00	35.07/ 9.33	36.00/ **6.00**	37.03/23.87	**42.59**/ 9.41	28.97/ **2.80**	58.00/12.14
Qwen2.5-VL-7B	26.90/ **5.00**	35.33/ 6.67	49.10/10.00	41.28/18.71	15.18/ 5.88	13.46/ 3.74	28.00/ 7.86

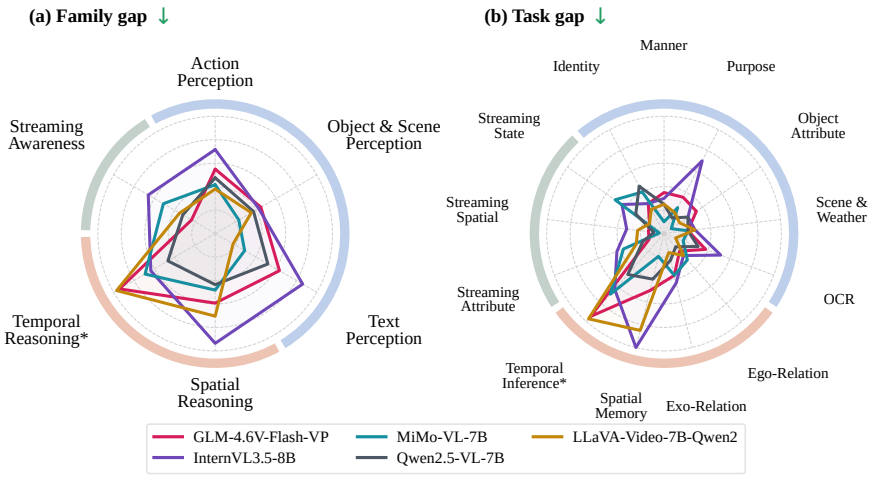

Figure 7: Radar-chart comparison across benchmark capabilities.* denotes Temporal Inference and Temporal Reasoning represented in own axis.

B DIAGNOSIS

B.1 DIAGNOSTIC LAYER AND HYPERPARAMETER SELECTION

For prefix calibration, we inspect how frame-attention profiles change as the target evidence moves across positions within the input window. We select the layer groups showing the strongest changes in attention profiles across evidence positions: layers 18-23 for GLM-4.6V-Flash and layers 7-9 for InternVL3-8B(Figures 10 and 11). Prefix calibration is applied only to these selected layers.

For each diagnostic intervention, we tune the intervention strength on an independently sampled 20% subset of Temporal Reasoning. We sweep $\{0, 0.25, 0.50, 0.75, 1.00\}$ for the prefix-calibration coefficient λ and the V-RoPE compression coefficient α, selecting the best-performing setting for each model and intervention. This yields $\lambda = 0.50/0.75$ and $\alpha = 0.25/0.50$ for InternVL3-8B / GLM-4.6V-Flash, respectively. Full sweep results are reported in Tables 8 and 9.

Table 7: Ablations of axis and compressing query-key for visual-RoPE compression of GLM-4.6V-Flash.

Model	**Setting**	α	**Acc. (%)**	**Range (pp)**
GLM-4.6V-Flash	Vanilla	1.00	35.25	21.61
GLM-4.6V-Flash	T-only	0.50	33.46	26.06
GLM-4.6V-Flash	H+W-only	0.50	35.10	17.55
GLM-4.6V-Flash	Always-compress	0.50	30.03	14.15
GLM-4.6V-Flash	All	0.50	36.96	15.46

Compressed conditions use $\alpha = 0.50$, a mean anchor, and apply only to text queries. All denotes all T/H/W axes. Ranges are in percentage points (pp). Invalid responses are retained and counted as incorrect.

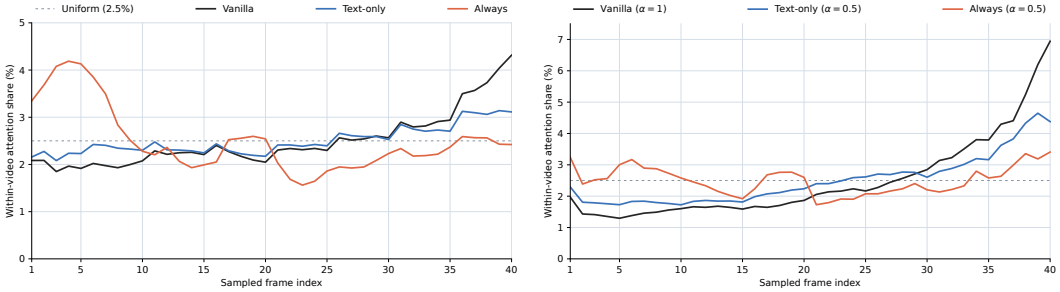

Figure 8: Attention distributions of GLM and InternVL3

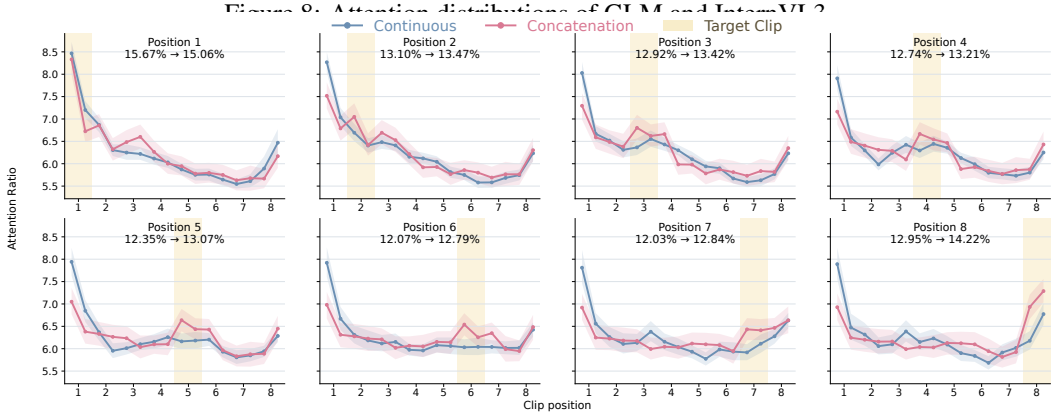

Figure 9: Attention under continuous and concatenated video contexts across all eight evidence positions, complementing the selected positions in Figure 2. Yellow bands mark the target clip; panel annotations report its attention share under continuous and concatenated contexts, respectively.

Table 8: Prefix calibration λ sweep on a 20% subset of Temporal Inference. Layer indices are zero-based. Mean is accuracy (%); Range is the accuracy range across evidence positions (percentage points). Best values within each model are bolded.

	GLM-4.6V-Flash Layers 18–23		InternVL3-8B Layers 7–9	
λ	Mean ↑	Range ↓	Mean ↑	Range ↓
0.00 (Vanilla)	24.76	28.85	**24.76**	16.35
0.25	25.36	32.69	24.16	16.35
0.50	**26.20**	27.88	24.16	**14.42**
0.75	25.00	**24.04**	24.04	16.35
1.00	24.52	25.96	23.92	**14.42**

Table 9: Visual-key RoPE compression α sweep on a 20% subset of Temporal Inference. Mean denotes accuracy (%); Range denotes the range of position accuracies (pp). Best values within each model are bolded.

	GLM-4.6V-Flash		InternVL3-8B	
α	Mean ↑	Range ↓	Mean ↑	Range ↓
0	22.62	33.33	42.48	26.90
0.25	25.00	**28.57**	**44.79**	**16.90**
0.50	**26.79**	38.10	43.89	17.14
0.75	23.81	**28.57**	41.70	26.67
1 (Vanilla)	25.00	**28.57**	40.11	**16.90**

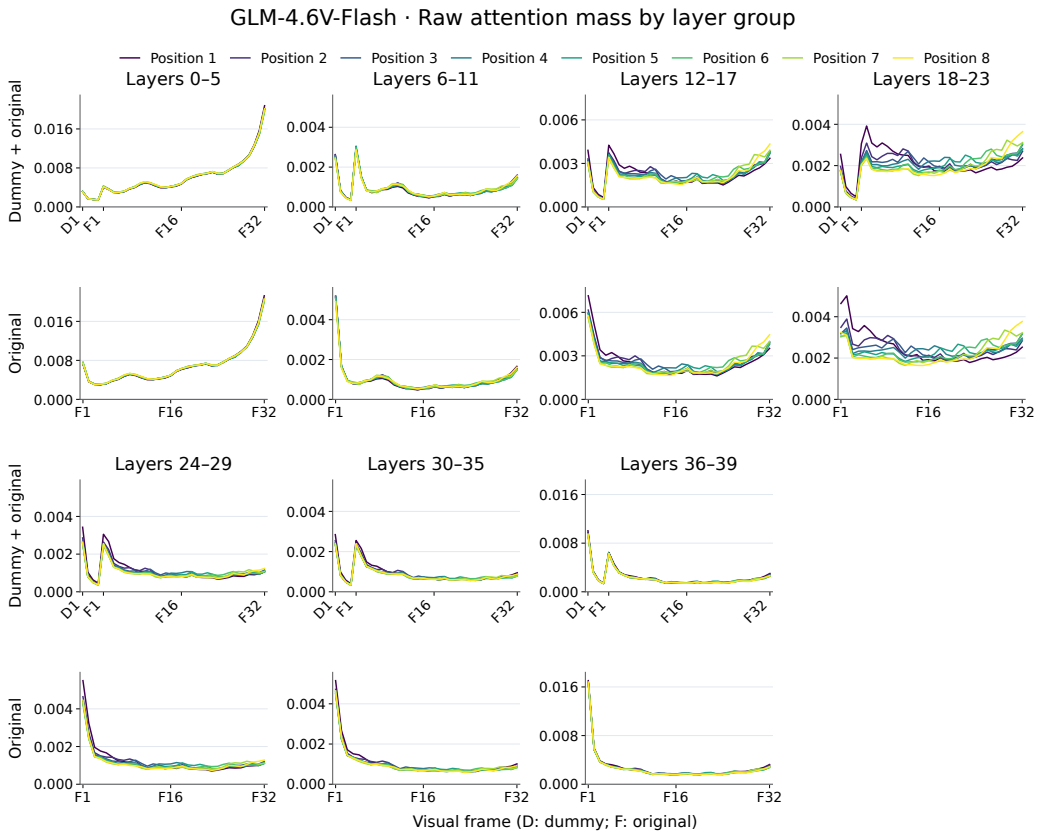

Figure 10: Frame-level attention profiles across GLM-4.6V-Flash layer groups for original inputs and inputs with a black-frame prefix. Colors indicate different evidence positions. Layers 18-23 exhibit pronounced changes in attention allocation as the evidence position varies.

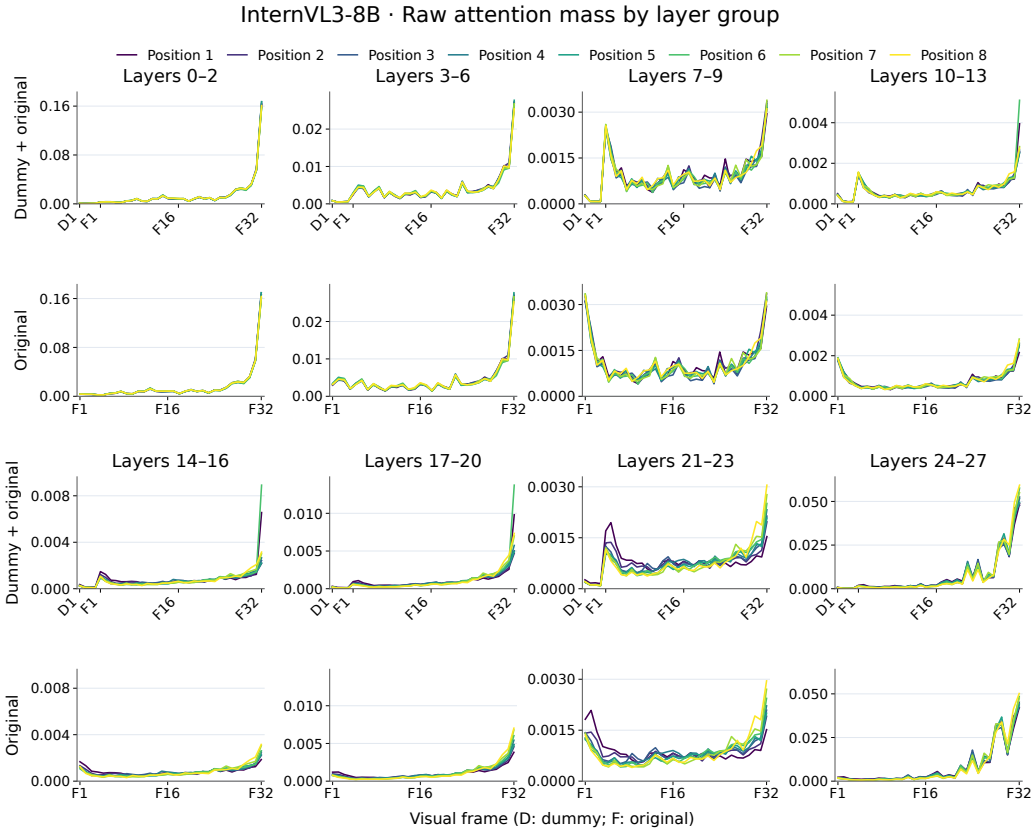

Figure 11: Frame-level attention profiles across InternVL3-8B layer groups for original inputs and inputs with a black-frame prefix. Colors indicate different evidence positions. Layers 7–9 exhibit pronounced changes in attention allocation as the evidence position varies. Late-layer groups are excluded from calibration because intervention in these layers causes severe answer-choice collapse.

C MITIGATION

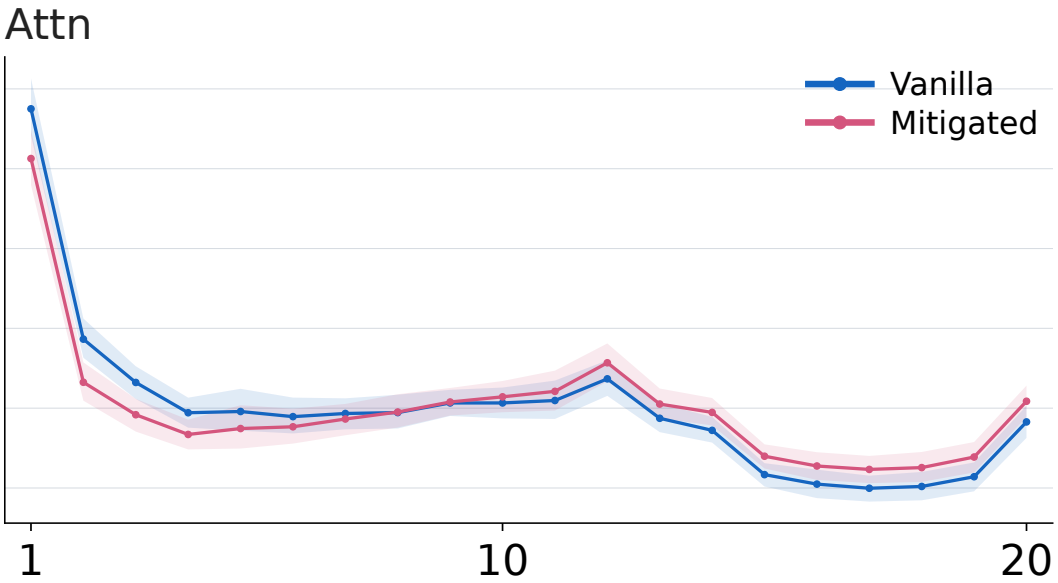

Figure 12: Attention profiles of the vanilla and fine-tuned models evaluated on Spatial Reasoning capability. Shaded bands indicate pointwise 95% percentile confidence intervals from 2,000 paired, task-stratified QID bootstrap resamples.

Table 10: Task-level comparison of continuous and concatenated training data. Each entry reports mean accuracy / positional range.

Task	Vanilla	Continuous	Concat.
Overall	38.91 / 31.49	61.88 / 17.06	54.64 / 20.98
Action Identity	42.69 / 30.22	57.06 / 13.53	53.67 / 21.61
Action Manner	30.48 / 19.90	39.85 / 19.82	42.42 / 19.85
Action Purpose	30.47 / 29.65	39.97 / 17.35	43.55 / 14.85
Object Attributes	47.85 / 25.46	56.93 / 17.28	54.54 / 19.24
Environmental Context	41.88 / 24.06	57.24 / 11.73	62.64 / 12.00
OCR	55.45 / 37.17	78.69 / 20.10	73.52 / 13.72
Egocentric Relation	25.58 / 26.70	45.28 / 14.78	42.75 / 19.47
Exocentric Relation	30.83 / 32.90	49.23 / 16.67	47.80 / 20.00
Spatial Memory	31.10 / 36.95	65.05 / 22.30	56.48 / 22.50
Time-conditioned Recognition	30.51 / 36.76	74.79 / 15.88	45.54 / 36.10