A Comprehensive Evaluation of Positional Bias in Multi-Video Summarization with Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) are increasingly used for video understanding, yet their reliability under multi-video inputs remains poorly understood. We present the first systematic study of positional bias in multi-video summarization, where summary quality varies with input order rather than underlying content. To support this analysis, we build an evaluation framework with three complementary metrics: Coverage, which measures how well summaries capture key information; Positional Bias, which quantifies directional preference for earlier or later inputs; and Positional Sensitivity, which captures how the quality of the summary for the same underlying video fluctuates across permutations. Evaluating diverse MLLMs on multi-topic datasets, we find that positional effects vary across domains and are further influenced by factors such as video duration and visual input configuration. These results establish a benchmark setting for diagnosing positional bias in multi-video summarization and highlight the need for more robust, order-invariant multimodal systems.

1 Introduction

Video summarization aims to condense video content into shorter yet informative representations while preserving essential information and temporal coherence (Kansal et al., 2023). As Multimodal Large Language Models (MLLMs) become increasingly capable of processing video inputs, they have also become a promising foundation for open-ended video summarization.

Recent MLLMs, such as Video-LLaMA (Zhang et al., 2023) and Gemini-2.5 Pro (Comanici et al., 2025), combine powerful language model backbones with visual encoders to handle text, images, and videos within a unified framework. Correspondingly, a growing body of benchmarks has been proposed to evaluate MLLMs across a range of abilities, including reasoning (Yue et al., 2024), long-video understanding (Zhou et al., 2025), temporal perception (Yu et al., 2019), and video summarization (Jung and Kim, 2025). Despite this progress, the behavior of MLLMs under *multi-video* inputs remains much less understood.

A particularly important concern is *positional bias*, where model performance changes with the order of inputs rather than with the underlying content itself. In natural language processing, large language models have been shown to exhibit systematic positional bias, such as favoring early content in news summarization (Grenander et al., 2019) and demonstrate U-shaped attention patterns that neglect middle content (Liu et al., 2024). These findings motivate a natural question for multimodal systems: **when multiple videos are provided together, do MLLMs summarize them differently depending on their input order?**

While prior studies have examined positional bias in multi-image understanding (Tian et al., 2025) and video QA through probing-based analyses (Xia et al., 2025), multi-video summarization has not yet been systematically studied from this perspective. This setting warrants separate investigation because, unlike QA-style tasks that often require locating specific evidence, summarization requires models to integrate and synthesize salient information across all input videos. In realistic scenarios, users may need to browse groups of related videos or summarize topic-based search results. If a model systematically favors certain positions, critical information may be overlooked simply because it appears earlier, later, or in a less favorable part of the sequence.

In this work, we present the first systematic study of positional bias in multi-video summarization. To support this study, we construct a benchmark based on ActivityNet (Heilbron et al., 2015) and the News Video Dataset (Whitehead et al., 2018), covering diverse domains and video durations, and build an evaluation framework tailored to this set-

ting. Because lexical-overlap metrics are insufficient for this highly abstractive task (as discussed in Appendix Section C.1), we adopt reference-based evaluation with LLM-as-a-Judge. Our framework uses three complementary metrics: **Coverage**, which measures whether a summary captures human-preferred key information; **Positional Bias**, a signed global index that reflects systematic preference for earlier or later positions; and **Positional Sensitivity**, which measures how the coverage of the same underlying video varies when it appears at different positions under order permutations.

Ideally, a fair multi-video summarization system should treat the same video consistently regardless of where it appears in the input sequence. However, our experiments show that current MLLMs remain strongly affected by input order. Importantly, these positional effects do not always manifest in the same way: some settings exhibit a clearer directional preference toward earlier or later inputs, while others show substantial per-video fluctuation across positions without a strong aggregate directional trend. Taken together, these findings suggest that low aggregate bias alone should not be interpreted as evidence of position invariance. Anonymous code and supplementary materials are available at `https://anonymous.4open.science/r/annoym07`.

Our contributions are as follows:

1. **Benchmark and evaluation setup**: We construct a multi-video summarization benchmark that enables controlled analysis under varying numbers of videos, video durations, and input configurations.

2. **Systematic positional-bias analysis**: We conduct a large-scale empirical study across both open-source and proprietary MLLMs, multiple domains, and diverse input settings. Our results show that positional effects in multi-video summarization are substantial and context-dependent.

3. **Preliminary mitigation exploration**: We examine prompt-based strategies as an initial attempt to reduce positional bias and discuss their implications for building more order-robust multi-video summarization systems.

2 Related Work

2.1 Positional Bias

While MLLMs have been extensively evaluated for task capability, much less is known about their robustness when multiple visual inputs must be summarized jointly. Positional bias refers to the tendency of models to favor certain input positions independent of the underlying content. In NLP, such effects have been documented in several settings. Large language models show lead bias in news summarization (Grenander et al., 2019) and the "lost in the middle" effect in long-context understanding (Liu et al., 2024). Position-dependent preferences have also been observed in LLM-as-a-Judge evaluations (Wang et al., 2024; Shi et al., 2025; Koo et al., 2024), recommendation settings (Hou et al., 2024), and long-form summarization faithfulness (Wan et al., 2025). Prior mitigation efforts include methods such as bootstrapping (Hou et al., 2024) and modified attention mechanisms (Zhang et al., 2024), although these approaches may introduce additional computational cost or may not transfer easily to closed-source models.

In multimodal settings, recent work has begun to identify analogous effects. Tian et al. (2025) show that reordering images can substantially affect reasoning outcomes in multi-image understanding. Video-LevelGauge (Xia et al., 2025) studies video LLMs through probing-based analyses that place relevant visual evidence at different positions. Our work differs in both task structure and evaluation target: rather than testing whether a model can retrieve localized evidence, we study whether it can maintain balanced coverage when summarizing multiple videos under different input orders.

2.2 Video Summarization and Automatic Evaluation Metrics

Video-to-text summarization has received increasing attention because natural language offers a convenient interface for consuming video content (Tsai et al., 2019). Recent work examines how MLLMs balance visual and textual information during summarization (Pennec et al., 2025) and explores QA-based protocols for evaluating temporal understanding in video summaries (Jung and Kim, 2025). However, prior work has not systematically studied positional bias in multi-video summarization.

Automatic evaluation for summarization has evolved across several paradigms. Early metrics such as BLEU (Papineni et al., 2002), ROUGE-

L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) focus on lexical overlap between generated and reference texts. While useful in some settings, such metrics are limited for highly abstractive summarization because they do not directly assess factual coverage, semantic adequacy, or coherence. Semantic metrics such as BERTScore (Zhang et al., 2020) partially address this issue by measuring similarity between contextualized embeddings, but they still do not directly determine whether a summary preserves the specific key information that human annotators regard as important.

Recent LLM-as-a-Judge methods (Fu et al., 2024; Liu et al., 2023) enable more flexible semantic evaluation by leveraging the language understanding abilities of large language models. Our work builds on this line of research, using reference-based evaluation to assess whether summaries preserve human-preferred key information across different input orders in multi-video summarization.

3 Evaluation

Figure 1 illustrates our evaluation pipeline. We construct multi-video inputs, prompt the model to generate aligned per-video summaries, and evaluate controlled order permutations under different model families, domains, video durations, frame counts, and resolutions. We then quantify positional effects using three metrics: *Coverage*, *Positional Bias*, and *Positional Sensitivity*.

3.1 Data Collection

We build our benchmark from two public datasets with human-written reference summaries, covering four scenario categories: *News*, *Sports*, *Daily Tutorial*, and *Multi*. For *News*, we use the News Video Dataset (Whitehead et al., 2018), which contains short YouTube news clips paired with concise summaries. For the remaining three categories, we sample videos from ActivityNet (Heilbron et al., 2015). Specifically, *Daily Tutorial* includes cooking and personal-care videos, while *Multi* combines videos from different categories.These scenarios reflect common multi-video consumption settings, such as news browsing, sports highlights, tutorial exploration, and mixed-topic recommendation feeds.

To study the interaction between positional effects and video duration, we partition inputs into short videos (approximately 1–2 minutes, used for all four scenarios) and long videos (approximately 3–5 minutes, used for *Sports* and *Daily Tutorial*). For each domain–duration configuration, we randomly sample 20 video groups, yielding 720 unique videos in total. Although the number of unique videos is limited, permutation-based evaluation produces thousands of order-controlled instances, enabling systematic analysis under a manageable annotation and inference budget.

3.2 Summary Generation and Alignment

Models. We evaluate both proprietary and open-source MLLMs. The proprietary models are GPT-4.1 (OpenAI, 2025) and Gemini-2.5 Pro (Comanici et al., 2025). The open-source models are VideoLLaMA3-7B (Zhang et al., 2023), InternVideo2.5-8B (Wang et al., 2025), InternVL2.5-26B (Chen et al., 2025), Qwen3-VL-30B-A3B-Instruct (abbreviated as Qwen3-VL-30B hereafter) (Bai et al., 2025), Phi-3.5-Vision-Instruct-4B (Abdin et al., 2024), and Gemma-3 27B-IT (Team et al., 2025).

Input scale. We focus on two input scales: pairwise inputs with two videos (P=2) and listwise inputs with four videos (P=4). For each scale, we use a cyclic permutation design so that each video appears at every position exactly once. This supports balanced within-video positional analysis while keeping the number of permutations tractable. We use domain-scale notation (e.g., Sports-2, News-4) to denote experiments on a specific domain with a given number of input videos.

Input configuration. Because model interfaces differ, we adapt visual input construction while keeping it fixed within each model across all permutations. Gemini 2.5 Pro supports direct video input, whereas GPT-4.1 receives frames extracted at 0.25 fps. For open-source models, we apply uniform frame sampling (4 or 8 frames per video), insert natural-language delimiters between videos, and resize frames to 384×384 (or 448×448 for the Intern family).

Per-video summary alignment. We use a two-stage alignment strategy. First, we instruct the model to produce per-video summaries with a fixed template such as `[video0] ...`, `[video1] ...`, and extract the segments by pattern matching. When the output deviates from the template, we apply a fallback semantic alignment: we split the reference summaries and model output into sentences, compute sentence-level semantic similarity, and assign each generated sentence to the video with the highest similarity score. Sentences assigned to the same video are then concatenated into

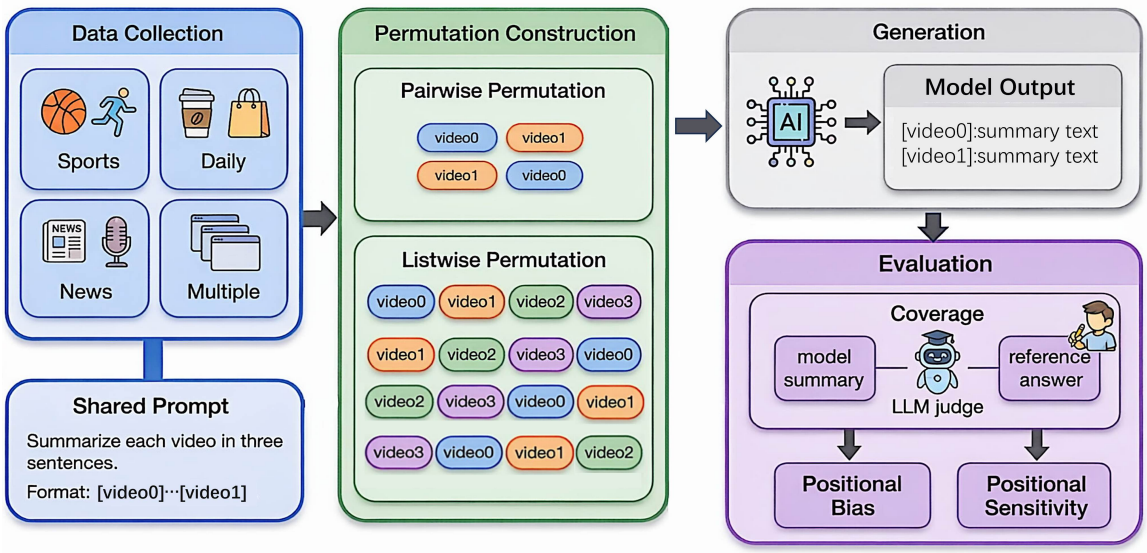


Figure 1: Overview of our evaluation pipeline. We quantify positional effects with three complementary metrics. **Coverage** measures how well each aligned per-video summary preserves key information. **Positional Bias** captures systematic preference for earlier versus later input positions through a signed, coverage-weighted index. **Positional Sensitivity** measures position invariance by tracking how the coverage of the same underlying video varies across order permutations.

the aligned per-video summary.

3.3 Evaluation Metrics

Let P denote the number of input videos in a multi-video instance, and let $p \in \{1, \ldots, P\}$ denote the position index of a video within that instance. Let N denote the total number of evaluated samples.

3.3.1 Coverage.

We first compare standard summarization metrics—BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020)—but find that they provide limited discrimination for positional bias in our multi-video setting. We therefore adopt *Coverage* as the primary signal, following the sentence-level evaluation spirit of PoSum-Bench (Sun et al., 2025).

Given a reference summary r_i segmented into sentences and a model-generated summary s_i, we use GPT-4o mini (OpenAI, 2024) as a judge to determine, for each reference sentence, whether it is semantically covered by s_i. Paraphrases and minor wording differences are allowed.

Let $N_{\text{miss}}^{(i)}$ denote the number of reference sentences judged as not covered, and let $N_{\text{ref}}^{(i)}$ denote the total number of reference sentences for sample i. We define:

$$\text{Coverage}_i = 1 - \frac{N_{\text{miss}}^{(i)}}{N_{\text{ref}}^{(i)}}. \tag{1}$$

Because the reference summaries are human-written and reflect salient content, Coverage provides a direct measure of alignment with human judgments of importance.

3.3.2 Positional Bias.

To quantify whether a model systematically favors earlier or later positions, we assign linear position weights:

$$\alpha_p = \frac{2p - P - 1}{P - 1}, \tag{2}$$

where $\alpha_1 = -1$ and $\alpha_P = 1$, yielding a normalized scale from -1 to $+1$ over the P positions in a multi-video input.

For each sample i, let $\text{Coverage}_{i,p}$ denote the coverage of the video appearing at position p. We define the per-sample bias index as:

$$\text{Bias}_i = \frac{\sum_{p=1}^{P} \alpha_p \cdot \text{Coverage}_{i,p}}{\sum_{p=1}^{P} \text{Coverage}_{i,p} + \epsilon}, \tag{3}$$

where ϵ is a small constant for numerical stability. The global bias index is then:

$$\text{Bias}_{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \text{Bias}_i. \tag{4}$$

Negative values indicate primacy bias, positive values indicate recency bias, and values near zero indicate weak aggregate directional preference.

3.3.3 Positional Sensitivity.

To quantify position invariance at the per-video level, we track the coverage of the same underlying video across positions under order permutations. For each sample i, let $c_{i,p}$ denote the coverage when the target video is placed at position p. We define:

$$\text{Positional Sensitivity} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\text{Var}_p(c_{i,p})}. \tag{5}$$

Lower Positional Sensitivity indicates stronger position invariance, whereas higher values indicate greater dependence on input position.

4 Experimental Results

Overall, our experiments show that positional bias in multi-video summarization are substantial, highly context-dependent, and not fully captured by aggregate bias alone. We analyze them from three perspectives: domain and model family under short-duration inputs, video duration for proprietary models, and visual input configurations for Qwen3-VL-30B.

4.1 Domain and Model Effects

We first examine the short-duration setting, where all models are evaluated across all domains. Figure 2 and Table 1 show that positional effects vary strongly across domains and models, rather than following a single universal pattern.

Positional bias vary markedly across domains. Multi-2 and Sports-2 show relatively consistent primacy-leaning behavior across many models, whereas Daily-2 is weaker and less consistent. News-2 is the most heterogeneous setting: Phi-3.5-Vision and Qwen3-VL-30B show strong recency, while VideoLLaMA3-8B and Gemma-3-27B-IT lean toward primacy. News tends to produce relatively high Positional Sensitivity for several models, especially in the pairwise setting, making it a useful stress test for position-invariant multi-video summarization. A plausible reason is that news videos are fact-dense, while their short reference summaries make failures to capture key information more heavily penalized in Coverage.

Aggregate bias alone is insufficient. GPT-4.1 on News-2 illustrates this clearly: its aggregate bias is 0.00, yet its Positional Sensitivity reaches 0.264, the highest value in that setting. One plausible explanation is that even within the same domain, positional behavior may vary across more fine-grained content types, causing different directional tendencies to cancel out in the aggregate while strong per-video sensitivity remains.

Positional behavior also varies across model families. The proprietary models, Gemini-2.5-Pro and GPT-4.1, usually have flatter curves and smaller aggregate bias magnitudes than most open-source models, but they are still far from position-invariant. For example, Gemini-2.5-Pro shows clear recency on News-4 ($+6.44\%$) with high sensitivity (0.252). Open-source models show larger dispersion and less stable positional strategies. VideoLLaMA3-8B is consistently primacy-biased across many settings, whereas Qwen3-VL-30B and Phi-3.5-Vision show sharper regime shifts; for example, Qwen3-VL-30B moves from strong recency on News-2 ($+21.95\%$) to slight primacy on News-4 (-2.50%).

4.2 Effect of Video Duration

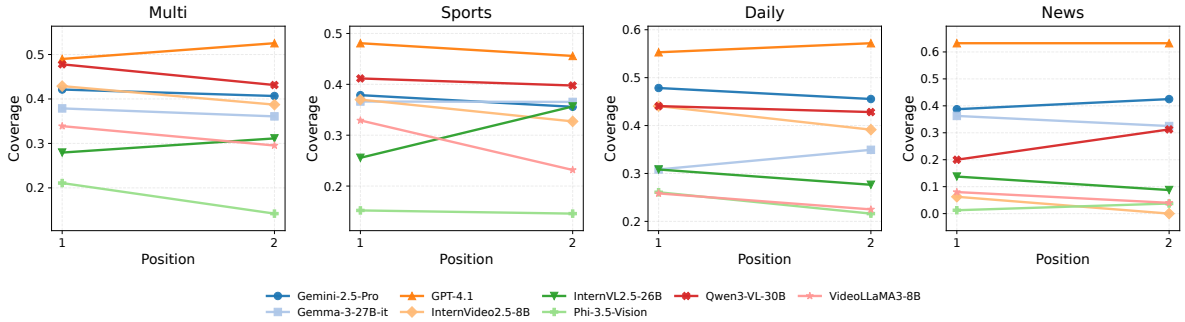
We next compare short- and long-duration videos for the proprietary models. As shown in Figure 3, longer videos generally reduce overall Coverage, consistent with the greater compression difficulty of longer inputs.

Duration effects depend on input scale. Under four-video inputs, the two proprietary models show relatively stable bias directions and similar Positional Sensitivity across durations. Sports-4 changes very little, whereas Daily-4 shows somewhat stronger duration dependence. This effect is most visible for Gemini-2.5-Pro, whose bias shifts from -1.14% under short-duration inputs to -5.14% under long-duration inputs. By contrast, under pairwise inputs, changes in duration can coincide with directional reversals.

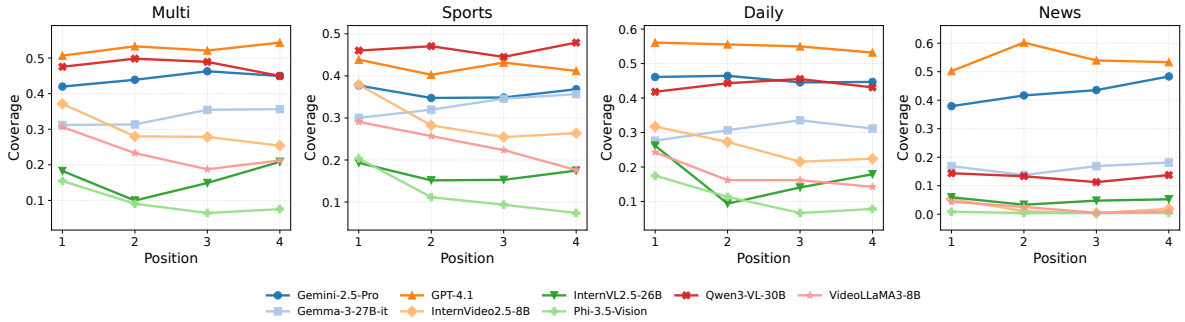
4.3 Effect of frame count and resolution

Finally, we study controlled ablations on Qwen3-VL-30B by varying frame count and input resolution.

Some patterns persist across visual configurations. In Multi-4, the curve remains broadly elevated around position 2, suggesting that this tendency is not solely an artifact of a single sampling configuration. Moreover, richer visual input is associated with improved position invariance in this ablation: under fixed resolution, Positional Sensi-



(a) Coverage versus position for pairwise inputs (P=2). Downward slopes indicate primacy bias; upward slopes indicate recency bias.



(b) Coverage versus position for listwise inputs (P=4).

Figure 2: Coverage curves across four domains under pairwise and listwise settings.

Model	Duration	Daily-2		Daily-4		Multi-2		Multi-4		News-2		News-4		Sports-2		Sports-4	
		Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity
Gemini-2.5-pro	Short	-2.45	0.073	-1.14	0.108	-1.76	0.083	2.14	0.127	4.62	0.143	6.44	0.252	-3.12	0.082	-0.58	0.117
Gemini-2.5-pro	Long	1.66	0.092	-5.14	0.123	–	–	–	–	–	–	–	–	-2.72	0.078	-0.59	0.115
Gemma-3-27B-IT	Short	6.23	0.115	3.64	0.127	-2.42	0.075	4.35	0.107	-5.45	0.198	3.49	0.199	-0.11	0.067	4.91	0.116
GPT-4.1	Short	1.68	0.096	-1.40	0.122	3.49	0.088	1.54	0.100	0.00	0.264	0.48	0.252	-2.67	0.087	-1.01	0.111
GPT-4.1	Long	-2.12	0.093	-0.98	0.109	–	–	–	–	–	–	–	–	3.85	0.079	-0.86	0.107
InternVideo2.5-8B	Short	-5.91	0.117	-10.93	0.154	-5.11	0.121	-9.96	0.187	-100.00	0.119	-41.46	0.102	-6.15	0.126	-10.55	0.186
InternVL2.5-26B	Short	-5.45	0.128	-9.98	0.174	5.39	0.164	6.59	0.155	-22.22	0.185	-1.23	0.154	16.49	0.123	-2.70	0.161
Phi-3.5-Vision	Short	-9.35	0.127	-25.95	0.129	-19.48	0.116	-22.81	0.117	50.00	0.079	-20.00	0.018	-2.10	0.125	-28.08	0.146
Qwen3-VL-30B	Short	-1.43	0.096	0.97	0.120	-5.12	0.073	-1.50	0.123	21.95	0.213	-2.50	0.175	-1.72	0.101	0.57	0.108
VideoLLaMA3-8B	Short	-6.90	0.158	-14.15	0.182	-6.87	0.146	-11.63	0.192	-33.33	0.118	-47.47	0.122	-17.38	0.140	-13.25	0.186

Table 1: **Coverage-based positional bias and sensitivity across domains and models.** Negative bias indicates primacy; positive bias indicates recency. Lower sensitivity indicates stronger position invariance.

tivity decreases from 0.123 at 8 frames per video to 0.101 at 16 frames per video.

5 Case Study

To complement the aggregate results, we analyze InternVideo2.5-8B, which exhibits a salient and reproducible failure pattern under multi-video generation. Across many four-video outputs, later summaries tend to reuse phrasing, sentence structure, or even event content introduced earlier in the same response. We refer to this cross-slot phenomenon as *textual carryover*.

5.1 Observation and Controls

When prompted to produce separate per-video summaries using explicit slot markers (e.g., `[video0]`, `[video1]`), InternVideo2.5-8B often fails to keep outputs independent across slots. Two symptoms recur: **template reuse**, where different slots share nearly identical syntactic scaffolds, and **cross-slot degradation**, where later slots often become partial or diluted rewrites of earlier ones. This behavior persists even when black separator frames are inserted between videos, suggesting that segmenta-

repetition_penalty	1vs2	2vs3	3vs4	avg
1.00	0.0798	0.0881	0.1250	0.0972
1.05	0.0070	0.0052	0.0092	0.0071
1.10	0.0000	0.0000	0.0000	0.0000

Table 2: **Adjacent cross-summary trigram overlap under different repetition penalties.** For each four-slot output (s_1, s_2, s_3, s_4), we compute $\mathrm{cov}(i{\to}i{+}1) = |\mathrm{ng}(s_i) \cap \mathrm{ng}(s_{i+1})|/|\mathrm{ng}(s_{i+1})|$ and report the mean across samples.

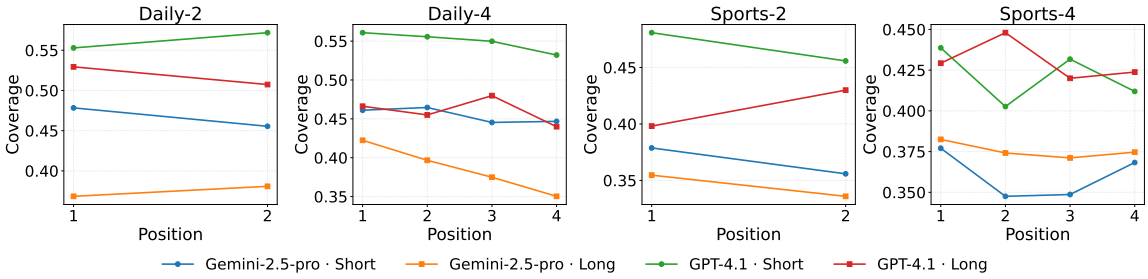
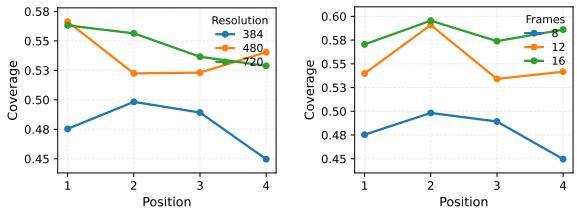


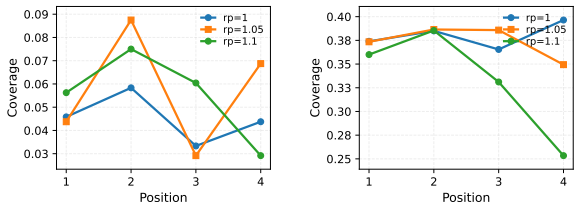
Figure 3: Coverage vs. position comparing short-duration (1–2 min) and long-duration (3–5 min) inputs for proprietary models.



(a) Varying resolution at fixed F=8.

(b) Varying frame count at fixed 384 × 384.

Figure 4: Coverage versus position for Qwen3-VL-30B on Multi-4.



(a) News-4.

(b) Sports-4.

Figure 5: Coverage versus position for InternVideo2.5-8B under different repetition penalties.

tion alone does not explain the failure.

We further use two controls to rule out simpler explanations. First, under a single-video setting, two highly similar hair-grooming videos still receive clearly differentiated descriptions, suggesting that scene similarity alone does not explain the templated cross-slot behavior. Second, when the same multi-event sequence is summarized as a single montage rather than as separate slot-based outputs, the regular slot-level templating pattern weakens, but repetition does not disappear. Together, these controls suggest that textual carry-over is neither solely driven by scene similarity nor purely an artifact of slot formatting; instead, slot-based prompting appears to amplify a broader generation pathology.

5.2 Failure Types

We distinguish two levels of failure. The examples below are drawn from summaries generated for the same video group under different input orders. In the more common **Level 1** case, the model still captures plausible video-specific details, but expresses them through nearly identical templates:

> `[video1]: The video shows` **`a large crowd gathered in front of a building`**`, with some people speaking to the camera...`
>
> `[video2]: The video shows` **`a large crowd marching in London`**`, with many people carrying signs and flags...`
>
> `[video3]: The video shows` **`a large crowd marching in the streets`**`, with many people carrying rainbow flags...`

In rarer but more severe **Level 2** failures, later slots become nearly identical to earlier ones (we observe one such case in the News domain even after inserting black separator frames):

> `[video2]: The video shows a large crowd marching down a street. The people are carrying flags and banners...`
>
> `[video3]: The video shows a large crowd marching down a street. The people are carrying flags and banners...`

Level 1 mainly reduces diversity and slot independence, whereas Level 2 directly harms content fidelity and Coverage.

5.3 A Decoding-Based Hypothesis

We interpret this failure pattern through a *generation-time coupling* hypothesis. Once a high-probability phrase or syntactic scaffold is produced early in the response, the growing textual prefix may bias later decoding toward reusing similar phrasing, structure, or even event content. This interpretation is consistent with prior observations

in autoregressive generation, where earlier generated tokens can shape later outputs (Zhang and Feng, 2022; Saito et al., 2025).

A qualitative example further supports this view. For the same set of input videos, changing only the input order can shift the opening template that then propagates to later slots. In one order, the first summary begins in an ongoing narrative style (e.g., "A group of people are..."), and the following summary adopts the same opening pattern. After swapping the order, the first summary instead starts with a more report-like template (e.g., "The video clip shows..."), and the next summary follows that style as well. This suggests that the first generated slot can act as a local stylistic anchor, increasing the likelihood of template reuse across later slots.

Table 2 provides supporting evidence for this view. Under default decoding, adjacent-slot trigram overlap is non-trivial and becomes larger for later slot pairs, consistent with progressive textual carryover within the same response. When the repetition penalty is increased, this overlap drops sharply and is nearly eliminated at higher values. This suggests that autoregressive decoding is a substantial contributor to the templated cross-slot reuse observed in InternVideo2.5-8B.

However, Figure 5 also shows that reducing textual overlap does not eliminate all positional effects. Position-dependent Coverage differences remain, and stronger repetition penalties can introduce new degradation, including weaker late-slot summaries and occasional slot omission. We therefore do not view decoding as a complete explanation. Instead, the evidence points to a **multi-causal** failure mode: decoding contributes materially to textual carryover, but the full positional behavior likely also depends on factors such as slot-structured prompting and competition among multiple video events during generation.

6 Mitigation Method

We explore a lightweight prompt-based intervention for mitigating positional bias by explicitly instructing the model to allocate balanced attention across input videos. We focus on VideoLLaMA3-7B, an open-source model that exhibits severe positional degradation in the baseline setting, and evaluate this intervention on $P{=}4$ inputs across two domains. **Prompting partially alleviates severe positional degradation.** As shown in Figure 6 and Table 3, the prompt intervention yields consistent

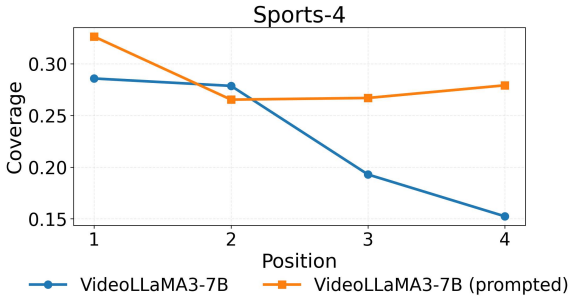


Figure 6: Coverage versus position for VideoLLaMA3-7B on Sports-4 with and without prompt intervention.

Setting	Daily-4		Sports-4	
	Bias	Sens.	Bias	Sens.
No Prompt	-17.06	0.171	-17.80	0.189
+Prompt	-13.06	0.169	-4.09	0.178

Table 3: Positional bias and sensitivity for VideoLLaMA3-7B with and without prompt intervention.

improvements for VideoLLaMA3-7B. On Sports-4, for example, the coverage curve becomes noticeably flatter, improving from a steep decline from 0.28 to 0.15 to a milder drop from 0.33 to 0.28 across positions. The aggregate bias magnitude is also substantially reduced. These results suggest that explicit instructions for balanced attention can mitigate severe position-dependent failures by narrowing performance gaps across positions, although they do not fully remove positional effects. At the same time, the persistence of template-like collapse in InternVideo under prompting indicates that prompt-based mitigation alone is unlikely to be sufficient; more robust solutions may require architectural or training-level changes.

7 Conclusion

We present the first systematic study of positional bias in multi-video summarization for MLLMs. Our results show that current models are not reliably position-invariant: summary quality can vary substantially with input order across domains, model families, and input settings. We further show that aggregate bias alone does not fully capture this phenomenon, as low global bias may still mask strong per-video positional sensitivity. Taken together, our benchmark, evaluation framework, and initial mitigation analysis provide a foundation for studying and improving order robustness in multi-video summarization.

8 Limitations

Input scale. Our main experiments focus on inputs of up to four videos. Positional effects under substantially longer input lists remain an open question.

Mechanistic understanding. We mainly provide a diagnostic analysis of positional bias. Although our case study offers initial clues about possible mechanisms, a more complete mechanistic account remains for future work.

Prompt scope. Our analysis is conducted under a fixed prompting protocol for controlled comparison. In real applications, users may provide more diverse prompts, and prior work has shown that language-side priors can sometimes dominate visual evidence. As a result, positional bias may vary under different prompt formulations.

9 Potential Risks

A potential risk of this work is that positional bias may be misused to place preferred videos in fixed positions, thereby shaping summary emphasis and causing overexposure of some content while reducing the visibility of others.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, and 1 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. *Preprint*, arXiv:2305.08845.

Woojun Jung and Junyeong Kim. 2025. QEVA: A reference-free evaluation metric for narrative video summarization with multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24632–24642, Suzhou, China. Association for Computational Linguistics.

Kajal Kansal, Nikita Kansal, Sreevaatsav Bavana, Bodla Krishna Vamshi, and Nidhi Goyal. 2023. A systematic study on video summarization: Approaches, challenges, and future directions. In *Proceedings of the 2nd Workshop on User-Centric Narrative Summarization of Long Videos*, NarSUM '23, page 65–73, New York, NY, USA. Association for Computing Machinery.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 517–545, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI Blog.

OpenAI. 2025. Introducing GPT-4.1 in the API. OpenAI Blog.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Galann Pennec, Zhengyuan Liu, Nicholas Asher, Philippe Muller, and Nancy F. Chen. 2025. Integrating video and text: A balanced approach to multimodal summary generation and evaluation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2403–2426, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Kuniaki Saito, Chen-Yu Lee, Kihyuk Sohn, and Yoshitaka Ushiku. 2025. Where is the answer? an empirical study of positional bias for parametric knowledge extraction in language model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1252–1269, Albuquerque, New Mexico. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Xu Sun, Lionel Delphin-Poulat, Christèle Tarnec, and Anastasia Shimorina. 2025. PoSum-bench: Benchmarking position bias in LLM-based conversational summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7985–8009, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10599–10609.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2025. On positional bias of faithfulness for long-form summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8791–8810, Albuquerque, New Mexico. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. 2025. Internvideo2.5: Empowering video mllms with long and rich context modeling. *Preprint*, arXiv:2501.12386.

Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. 2018. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001, Brussels, Belgium. Association for Computational Linguistics.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. 2025. Video-levelgauge: Investigating contextual positional

bias in large video language models. *Preprint*, arXiv:2508.19650.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9127–9134.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2022. Reducing position bias in simultaneous machine translation with length-aware framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6775–6788, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701.

A Prompts and Input Formatting

A.1 Per-video Summarization Prompts

We use a fixed per-video output structure to support reliable slot extraction and position-wise analysis. All models are instructed to return one summary per video in an explicit tagged format. For ActivityNet-based domains, we request three sentences per video for short-video settings and five sentences per video for long-video settings. For the News domain, whose reference summaries are much shorter (on average about 1.2 sentences), we reduce the requested number of sentences to one to better match the granularity of the reference summaries. We acknowledge that such fixed sentence budgets may constrain the maximum attainable coverage, however, this design helps standardize output granularity across models and supports more reliable position-wise comparison.

```
Summarize each video in {num_sentences}
sentence(s).
Format:
video0
<summary>
video1
<summary>
...
```

Input packaging. We package the input as tagged frame lists, where the frames under each `videoX` tag are arranged in chronological order.

```
Frames from video0: <frames>
Frames from video1: <frames>
...
```

A.2 Prompting with Black Separator Frames

To improve clip-boundary recognition for InternVideo2.5 in multi-video inputs, we optionally insert a black separator frame between consecutive clips.

```
You are shown a montage composed of several
separate video clips.
The montage contains {num_videos} clips
presented in one continuous sequence.
The input below is organized as text and frames:
- Before each clip, there is a text tag such as
'video0', 'video1', etc.
- Under each videoX tag, frames are listed
in strict chronological order as 'FrameK:
<image>'.
- A BLACK SEPARATOR frame '(black separator)'
appears between two clips and marks their
boundary.
- Some additional BLACK frames may appear at the
end as '(padding black)' for internal alignment.
Your task is to summarize each clip
independently based on its own frames.
```

A.3 LLM-as-a-Judge Prompt for Sentence-level Coverage

We evaluate Coverage using an LLM-as-a-Judge protocol. Given a reference summary split into sentences and a model-generated summary for the same video, the judge counts how many reference sentences are not covered by the model output. The prompt is shown below. In practice, we provide the judge with the full reference answer instead of feeding reference sentences independently. This design is motivated by the observation that reference summaries often contain cross-sentence dependencies, including anaphoric expressions and context carried over from preceding sentences. Judging sentences in isolation may therefore create ambiguity and reduce the reliability of coverage assessment.

```
You are given a reference summary for a video,
split into sentences,
and a model-generated summary for the same
video.

Your task:
1. For each reference sentence, decide whether
it is clearly covered by the model summary.
- A reference sentence is considered covered
if the model summary explicitly or implicitly
describes the same fact or event.
- The wording can be different, but the meaning
must match.
2. Count how many reference sentences are NOT
covered by the model summary.
3. IMPORTANT: Output ONLY a single integer (0,
1, 2, ...) with no explanation and no extra
text.

Reference sentences:
{numbered_refs}

Model summary:
{model_summary}

Now output the number of reference sentences
that are NOT covered by the model summary:
```

B Implementation Details and Default Input Configurations

B.1 Implementation Details

Table 4 summarizes the default input configurations used in our main experiments. These settings balance inference cost, model-specific constraints, and protocol consistency across model families. For LLM-as-a-Judge evaluation, we use `GPT-4o-mini`. For `InternVL2.5-26B` and `Qwen3-VL-30B`, INT8 quantization is applied during inference to reduce GPU memory usage. The full inference and evaluation pipeline was run on a single NVIDIA A800

Model	Size	Frames / Video	Resolution
Gemini-2.5-Pro	–	original	original
GPT-4.1	–	fps=0.25	original
VideoLLaMA3	7B	8	384×384
InternVideo2.5	8B	8	448×448
InternVL2.5	26B	8	448×448
Qwen3-VL	30B	8	384×384
Phi-3.5-Vision-Instruct	4B	8	384×384
Gemma-3	27B	4	384×384

Table 4: **Default input configurations.** For closed-source models, we follow their native preprocessing and sampling behavior. For open-source models, we fix the number of sampled frames and spatial resolution to balance inference cost and model-specific constraints.

80GB GPU, with a total runtime of approximately 300 hours for the main experiments.

B.2 Dataset Licenses and Usage Conditions

Our benchmark is constructed from two publicly available third-party datasets derived from web videos: the News Video Dataset and ActivityNet. For the News Video Dataset, the official repository states that the dataset is released for educational and research purposes only, while the videos remain protected under their respective YouTube licenses and belong to their original copyright holders. For ActivityNet, we rely on the official release of the dataset and the access channels provided by the benchmark creators.

C More Experimental Details

C.1 Lexical Metrics

Representative metrics. We report **BERTScore** and **ROUGE-L** as representative semantic and lexical metrics, respectively. Figure 7 shows that traditional automatic metrics are highly correlated with each other, with pairwise Spearman correlations mostly between 0.63 and 0.78, indicating substantial redundancy in model ranking.For lexical metrics, we use public implementations of BLEU, ROUGE, METEOR, and BERTScore. These metrics are computed with their default settings. By contrast, *Coverage* is only moderately correlated with these metrics (0.42–0.55), suggesting that it captures related but distinct aspects of summary quality. Tables 5 and 6 further show that BERTScore and ROUGE-L provide weaker discrimination of positional effects across domains and models. We therefore use Coverage as the primary metric in the main positional-bias analysis.

C.2 Additional Results

Figure 9 provides additional results that complement the analyses in the main text. The top row shows prompt-based mitigation results for VideoLLaMA3-7B on Daily-4, Multi-4, and News-4. The bottom row shows additional News-4 ablations for Qwen3-VL-30B under different visual input configurations, including varying frame count at fixed resolution and varying resolution at fixed frame count.

C.3 Cross-judge Robustness

Figure 8 provides an additional analysis of judge robustness on Sports-4. We compare the Coverage results of Gemini and GPT outputs when scored by two different LLM judges, GPT-4o-mini and Gemini-2.5-Flash.

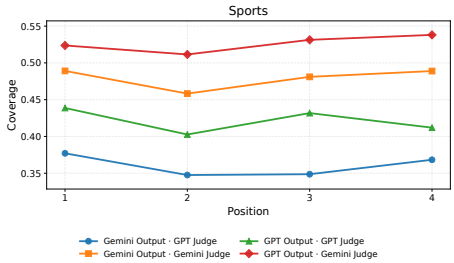


Figure 8: **Coverage under different judges on Sports-4.**

Although GPT-4o-mini is consistently stricter and assigns lower absolute Coverage scores than Gemini-2.5-Flash, the relative ordering between the compared models remains unchanged: GPT outputs are consistently scored higher than Gemini outputs under both judges at every position. In addition, the positional curves are broadly aligned across judges, including similar local patterns such as the dip around Position 2 followed by recovery. These results suggest that, while the absolute scale of Coverage depends on the judge model, the comparative conclusions are reasonably robust to judge choice.

D Examples

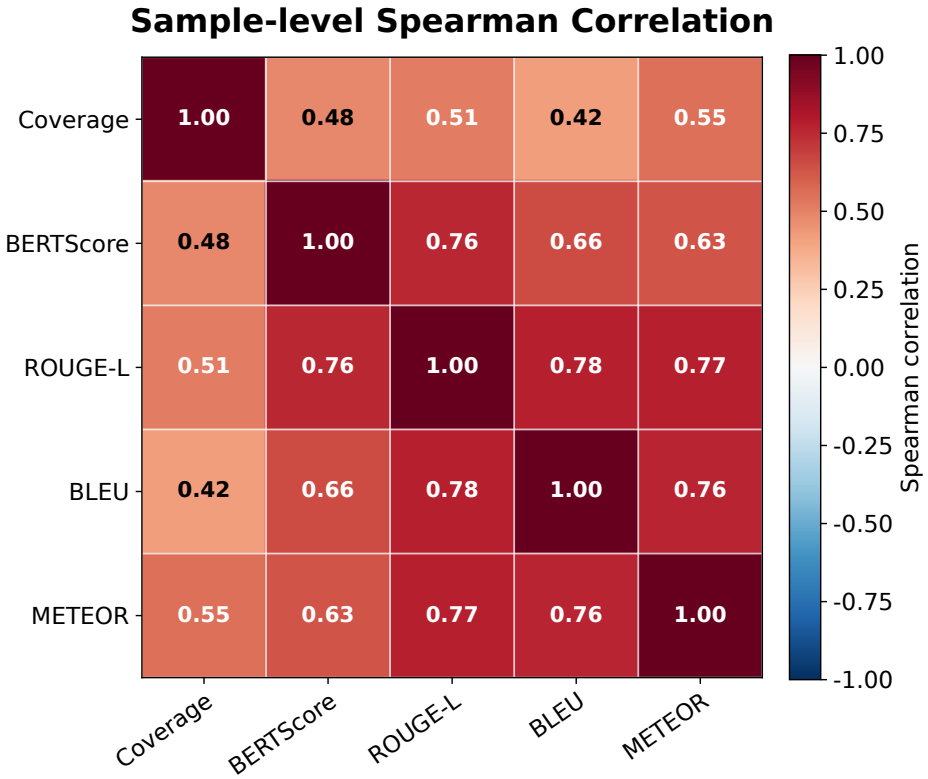


Figure 7: **Spearman correlation heatmap across evaluation metrics.** Correlations are computed from aggregated position-aligned observations over all runs. Traditional lexical metrics are highly correlated with each other, while Coverage remains positively correlated with them but is more sensitive to positional allocation effects.

Model	Duration	Daily-2		Daily-4		Multi-2		Multi-4		News-2		News-4		Sports-2		Sports-4	
		Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity
Gemini-2.5-pro	Short	-0.09	0.005	-0.07	0.006	-0.03	0.005	-0.00	0.005	-0.11	0.006	0.05	0.009	0.03	0.004	-0.05	0.006
Gemini-2.5-pro	Long	-0.03	0.004	-0.06	0.006	–	–	–	–	–	–	–	–	-0.01	0.004	-0.06	0.005
Gemma-3-27b-it	Short	0.02	0.005	-0.05	0.005	-0.05	0.005	0.01	0.005	-0.04	0.006	0.06	0.007	0.12	0.004	0.03	0.006
GPT-4.1	Short	-0.05	0.004	-0.12	0.006	-0.05	0.005	-0.09	0.006	-0.08	0.010	-0.00	0.009	0.07	0.005	-0.07	0.006
GPT-4.1	Long	-0.13	0.005	-0.09	0.006	–	–	–	–	–	–	–	–	-0.04	0.005	-0.12	0.005
InternVideo2.5-8b	Short	-2.84	0.105	-6.81	0.215	-1.46	0.077	-7.70	0.258	-13.86	0.212	-20.79	0.345	-5.92	0.150	-7.46	0.246
InternVL2.5-26b	Short	-3.64	0.118	-4.73	0.217	-6.82	0.164	-4.03	0.217	-2.83	0.095	-7.19	0.255	-1.36	0.074	-2.86	0.177
Phi-3.5-Vision	Short	-0.17	0.008	-0.21	0.014	-0.47	0.010	-1.23	0.086	-0.17	0.008	-0.24	0.010	-0.18	0.008	-0.18	0.014
Qwen3-VL-30b	Short	0.10	0.005	0.18	0.007	0.11	0.004	0.05	0.007	-0.07	0.009	0.02	0.006	0.17	0.005	0.11	0.006
VideoLLaMA3-8B	Short	-0.43	0.015	-2.13	0.126	-0.10	0.010	-2.84	0.106	-0.53	0.019	-0.40	0.021	-0.16	0.012	-1.96	0.126

Table 5: **BERTScore positional bias and sensitivity.** $\text{Bias}_{\text{global}}$ (shown as %) and Sensitivity are computed in the same way as in Table 1. The results show weaker discrimination of positional effects than Coverage.

Model	Duration	Daily-2		Daily-4		Multi-2		Multi-4		News-2		News-4		Sports-2		Sports-4	
		Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity	Bias (%)	Sensitivity
Gemini-2.5-pro	Short	-1.14	0.025	0.01	0.027	-0.08	0.023	0.89	0.026	-3.93	0.052	0.54	0.060	-3.52	0.018	-0.21	0.026
Gemini-2.5-pro	Long	0.72	0.023	-0.12	0.029	–	–	–	–	–	–	–	–	0.71	0.017	-0.69	0.023
Gemma-3-27B-IT	Short	2.49	0.022	-0.04	0.030	-2.08	0.022	0.28	0.027	-1.13	0.036	0.04	0.036	0.62	0.021	0.39	0.027
GPT-4.1	Short	0.79	0.027	-1.33	0.030	1.03	0.025	-0.60	0.027	-0.84	0.052	0.15	0.054	1.79	0.022	-0.62	0.026
GPT-4.1	Long	2.45	0.022	-1.02	0.028	–	–	–	–	–	–	–	–	-1.65	0.022	-1.27	0.025
InternVideo2.5-8B	Short	-1.16	0.042	-7.58	0.078	-0.43	0.054	-5.30	0.087	-4.61	0.054	-14.07	0.054	-1.91	0.056	-6.64	0.083
InternVL2.5-26B	Short	-5.71	0.038	-3.71	0.061	-5.31	0.044	-4.18	0.062	-5.89	0.035	-3.63	0.038	-1.13	0.029	-2.81	0.052
Phi-3.5-Vision	Short	-2.20	0.040	-1.89	0.050	-5.65	0.035	-6.00	0.059	-3.07	0.034	-4.58	0.037	-1.30	0.035	-0.12	0.043
Qwen3-VL-30B	Short	2.95	0.027	2.05	0.029	1.12	0.021	1.57	0.036	-2.53	0.042	1.34	0.039	3.77	0.026	1.98	0.029
VideoLLaMA3-8B	Short	-3.04	0.039	-5.56	0.074	-1.73	0.053	-5.17	0.075	-9.15	0.090	-12.71	0.111	-2.76	0.052	-6.30	0.070

Table 6: **ROUGE-L positional bias and sensitivity.** $\text{Bias}_{\text{global}}$ (shown as %) and Sensitivity are computed in the same way as in Table 1. Similar to BERTScore, ROUGE-L provides weaker sensitivity to positional effects than Coverage.

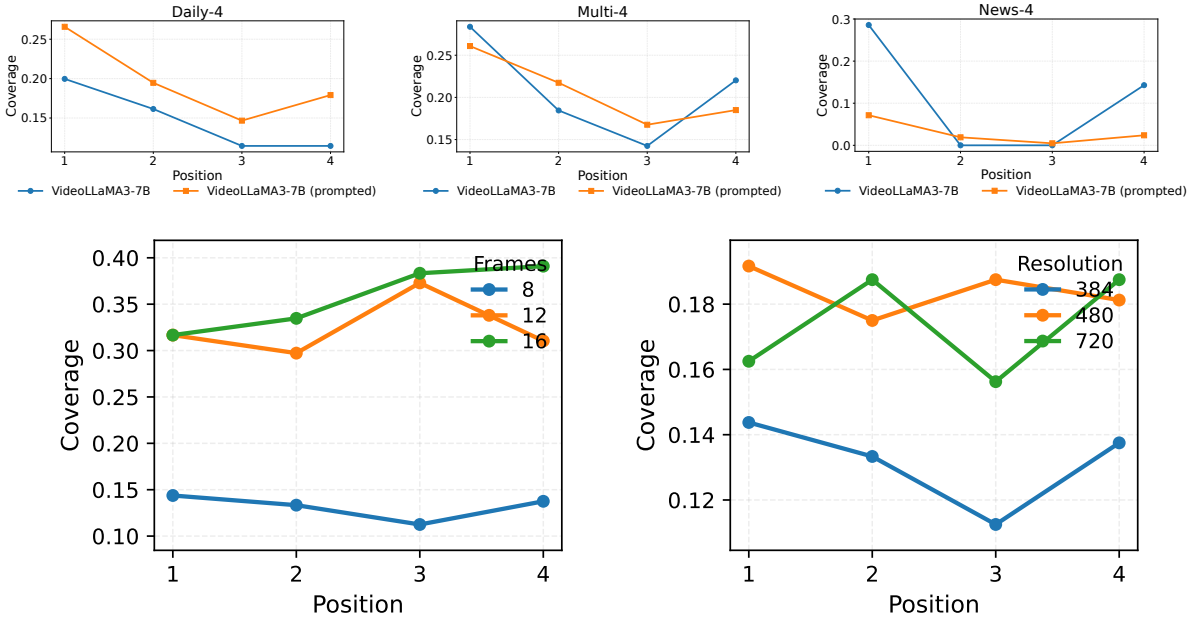


Figure 9: Additional positional-bias analyses. Top: Coverage versus position under prompt-based mitigation for VideoLLaMA3-7B on Daily-4, Multi-4, and News-4. Bottom: additional News-4 visual-input ablations for Qwen3-VL-30B. Left: varying frame count with resolution fixed at 384×384. Right: varying resolution with frame count fixed at 8.

Example 1.

Position 1

A man demonstrates how to start a fire in a campground fire pit using a ✓ fire starter. He strikes the rod several times over a pile of tinder, which includes wood shavings and lichen, creating sparks. After a few tries, the tinder ✓ catches fire, and he successfully starts a small flame while his ✓ children watch.

Position 2

A man at a campsite is preparing to start a fire inside a metal fire ring. He uses a ✓ fire striker to create sparks on a tinder bundle made of wood shavings and some white material. After a few tries, the tinder ✓ catches fire, and he successfully starts a small flame.

Position 3

A man demonstrates how to start a campfire using a ✓ fire starter in a metal fire pit. He scrapes a ferro rod over a tinder bundle of wood shavings and lichen, creating sparks to ignite it. After a few strikes, ✓ the sparks catch, and he successfully starts a small flame.

Position 4

A man at a campsite demonstrates how to start a fire using a ✓ ferro rod and tinder. He carefully arranges wood shavings and other flammable materials in a fire pit before striking the rod to create sparks. After a few tries, the sparks successfully ignite the tinder, and ✓ a small flame begins to grow.

Figure 10: **T**he four responses consistently preserve the core ignition-related details (e.g., *fire starter* and *catches fire*), while the contextual detail about children watching appears only in Position 1.

Example 2.

Position 1

A male weightlifter wearing a USA singlet performs a ✓ snatch lift in a competition setting. He successfully lifts the heavy barbell from the floor to an ✓ overhead position with explosive power and good form. After holding the lift, he drops the weight and celebrates with an ✓ enthusiastic jump.

Position 2

This video features weightlifter Kendrick Farris competing in a USA Weightlifting event. He executes a powerful and successful ✓ snatch, lifting a 157 kg barbell ✓ over his head in a single fluid motion. After holding the lift to secure the attempt, he drops the weight and celebrates with an ✓ exuberant jump.

Position 3

At a USA Weightlifting competition, an athlete successfully executes a ✓ 157kg snatch lift. After completing the lift and dropping the barbell, he celebrates his achievement with incredible enthusiasm. The celebration concludes with an impressively high and acrobatic backflip.

Position 4

This clip shows a male weightlifter competing in a USA Weightlifting event. He successfully completes a ✓ snatch lift, lifting a 157 kg barbell from the floor to ✓ overhead in one powerful motion. After holding the weight to confirm the lift, he drops the barbell and celebrates with an ✓ excited jump.

Figure 11: Across the four positions, the responses consistently preserve the core lifting action. However, the excited post-lift celebration is described differently in Position 3 as a backflip rather than a jump. Position 3 also does not explicitly mention the overhead phase.