A SYSTEMATIC EVALUATION OF POSITIONAL BIAS IN MULTI-VIDEO SUMMARIZATION WITH MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are increasingly used for video understanding, yet their reliability under multi-video inputs remains poorly understood. We study positional bias in multi-video summarization, where the quality of a per-video summary can change with the video's input slot even when the underlying content is unchanged. We construct a benchmark from ActivityNet and News videos, covering Cooking, Domestic, Leisure, and News settings with two- and four-video inputs. We evaluate nine open-source and proprietary MLLMs and measure position effects with three complementary metrics: Coverage, Directional Positional Bias (DPB), and Middle-Edge Gap (MEG). Our results show that positional effects are domain- and model-dependent: signed directional bias can be small even when middle positions underperform, and increasing visual or generation budgets does not uniformly remove the imbalance. We further analyze prompt-level mitigation methods. Together, the results show that multi-video summarization remains sensitive to input protocol and position, motivating 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 become a promising foundation for open-ended video summarization. Recent benchmarks have evaluated MLLMs on long-video understanding Zhou et al. (2025) and video summarization Jung & Kim (2025). Meanwhile, emerging *multi-video* benchmarks have shifted attention to inputs containing multiple videos, but they primarily target understanding, perception, or reasoning rather than summarization Peng et al. (2025); Bai et al. (2026). Less is known about *multi-video summarization*, where the model must generate one aligned summary for each video in the same input.

A key concern is *positional bias*, where model performance changes with input order rather than the underlying content. In NLP, large language models have been shown to favor early content in news summarization Grenander et al. (2019) and to exhibit "lost in the middle" behavior in long-context settings Liu et al. (2024). Recent multimodal work has also examined positional effects in multi-image understanding Tian et al. (2025) and video probing benchmarks Xia et al. (2025). Yet these protocols mainly test whether models can retrieve or use target evidence after its position changes. Multi-video summarization poses a different challenge: the model must bind each summary to the correct video while balancing limited summary detail across multiple input slots.

In this work, we present a systematic study of positional bias in multi-video summarization. We construct a benchmark from ActivityNet Heilbron et al. (2015) and the News Video Dataset Whitehead et al. (2018), covering four scenario categories: *Cooking*, *Domestic*, *Leisure*, and *News*. We evaluate both pairwise and four-video inputs using cyclic orderings, so that each video appears at each position exactly once. Because lexical-overlap metrics are insufficient for this highly abstractive task, we use a dataset-adaptive, reference-based Coverage measure: ActivityNet-derived domains are scored with an LLM-as-a-Judge protocol, while News uses extractive reference-fragment coverage. We then report Coverage, Directional Positional Bias (DPB), and Middle-Edge Gap (MEG) to distinguish overall information preservation, earlier-versus-later preference, and middle-position weakness. Anonymous code and data are available at `https://anonymous.4open.science/r/annoym07`.

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.

Our contributions are as follows:

1. **Benchmark and evaluation protocol**: We construct a benchmark over *Cooking*, *Domestic*, *Leisure*, and *News*. We also introduce an evaluation protocol that measures Coverage, Directional Positional Bias (DPB), and Middle-Edge Gap (MEG), separating earlier-versus-later preference from middle-position weakness.

2. **Model- and domain-dependent positional effects**: We evaluate nine open-source and proprietary MLLMs and show that positional effects are model- and domain-dependent. Several settings have near-zero DPB but negative MEG, which would be missed by directional bias alone.

3. **Robustness checks and mitigation analysis**: We examine visual budget, requested summary length, boundary format, and prompt placement, and evaluate prompt-level mitigation strategies.

2 Related work

2.1 Positional bias

Positional bias refers to models' sensitivity to where information or candidates appear in the input, beyond their underlying content. In text settings, prior work has studied lead or position bias in summarization Grenander et al. (2019); Schilcher et al. (2025), the "lost in the middle" effect in long-context QA and retrieval Liu et al. (2024), and position-dependent faithfulness in long-form summarization Wan et al. (2025). Similar order effects also appear when LLMs are used as judges, rankers, or recommenders Wang et al. (2024); Shi et al. (2025); Koo et al. (2024); Hou et al. (2024). These studies motivate a range of mitigation strategies, from training or attention-based methods to lightweight inference-time controls such as order perturbation and attention-guiding prompts Wang et al. (2024); Zhang et al. (2024); Wan et al. (2025); Tian et al. (2025).

In multimodal settings, recent work has begun to identify analogous effects. Tian et al. (2025) show that reordering images can substantially affect multi-image reasoning, while Video-LevelGauge Xia et al. (2025) studies video LLMs through probing tasks that place relevant visual evidence at different positions. Our work differs in both task structure and evaluation target. Rather than testing whether a model retrieves localized evidence from a moved probe, we study whether it maintains balanced information coverage when generating aligned summaries for multiple videos under different input orders.

2.2 Automatic evaluation for summarization

Automatic summarization evaluation has long relied on lexical-overlap metrics such as BLEU, ROUGE, and METEOR Chang et al. (2024); Papineni et al. (2002); Lin (2004); Banerjee & Lavie (2005), and later embedding-based metrics such as BERTScore Zhang et al. (2020). These metrics are inexpensive and scalable, but they mainly measure text similarity and do not directly assess whether the generated summary preserves or aligns with specific information in the reference.

Recent work has explored more semantic evaluation protocols. LLM-based evaluators such as GPTScore and G-Eval use large language models to judge generated text along flexible task-specific criteria Fu et al. (2024); Liu et al. (2023). QA-based metrics such as QuestEval evaluate summarization through question generation and question answering Rebuffel et al. (2021), and recent video-summary evaluation further uses multimodal QA to assess coverage, factuality, and chronology Jung & Kim (2025). These approaches provide richer diagnostics than surface overlap, but they can require multiple model calls per summary and are costly for large-scale position-wise evaluation over many videos, orders, and slots.

For short news summaries, extractive-fragment analysis provides another useful perspective by measuring partial phrase-level overlap Grusky et al. (2018). This line of work is especially relevant when references are concise and contain fixed named entities or event phrases.

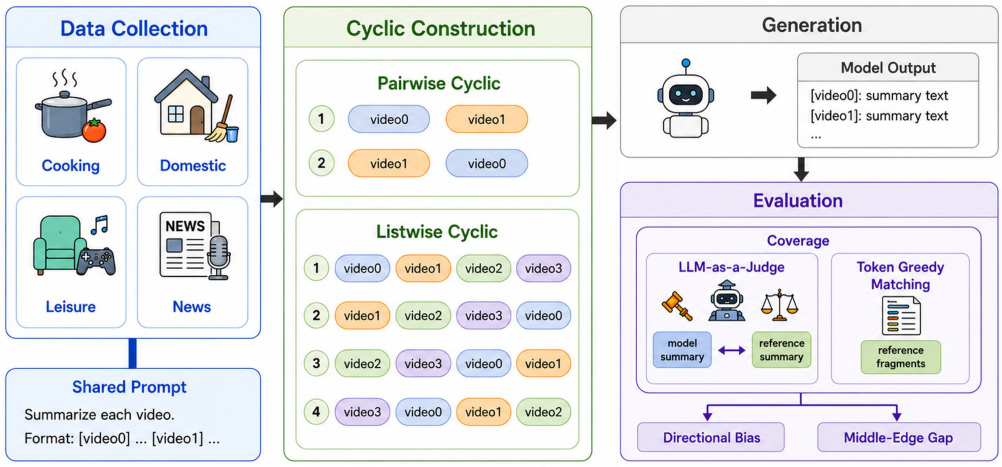


Figure 1: Overview of our evaluation pipeline. We construct cyclic orderings for pairwise and listwise multi-video inputs, generate aligned per-video summaries, and evaluate positional effects with three complementary metrics.

3 Evaluation

Figure 1 illustrates our evaluation pipeline. We construct multi-video inputs, prompt the model to generate aligned per-video summaries, and evaluate positional effects with cyclic orderings across multiple settings, including model families, domains, video durations, and visual-budget variants. We then quantify positional effects using three metrics: *Coverage*, *Directional Positional Bias* (DPB), and *Middle-Edge Gap* (MEG).

3.1 Data Collection

We build our benchmark from two public datasets with human-written reference summaries, covering four scenario categories: *Cooking*, *Domestic*, *Leisure*, and *News*. For *News*, we use the News Video Dataset Whitehead et al. (2018). For the other three categories, we sample from ActivityNet Heilbron et al. (2015): *Cooking* corresponds to eating and cooking activities, *Domestic* to household activities, and *Leisure* to social, relaxation, and recreational activities.

These categories instantiate common online video-use settings. Recent studies show that short-video platforms are used both functionally for learning and recreationally for entertainment, while social platforms also serve as important channels for news consumption Liang & Wang (2026); Altay et al. (2025).

To study the interaction between positional effects and video duration, we partition inputs into short videos (0–1 minute) and longer-duration videos (1–2 minutes). *Cooking* and *News* contain only short-video settings, while *Domestic* and *Leisure* contain both short- and longer-duration settings. For each domain–duration configuration, we randomly sample 27 video groups. The four-video groups are non-overlapping within each topic–duration configuration, while the pairwise and four-video settings may share videos within the same topic. This yields 708 unique videos in total. Although the number of unique videos is limited, cyclic evaluation produces thousands of order-controlled instances, enabling systematic analysis under a manageable inference budget.

3.2 Summary Generation and Alignment

Models. We evaluate both proprietary and open-source MLLMs. The main comparison includes InternVL3.5-8B, InternVL3.5-14B Wang et al. (2025), Qwen3-VL-8B, Qwen3-VL-30B-A3B Bai et al. (2025), MiniCPM-o-4.5 Cui et al. (2026), GLM-4.1V-9B-Thinking Team et al. (2026), MiMo-VL-7B-RL Xiaomi (2025), Gemini-3.1-Pro Google DeepMind (2026), and GPT-5.4 OpenAI (2026).

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., Cooking-2, News-4) to denote experiments on a specific domain with a given number of input videos.

Input configuration. All main runs use the same task protocol: the task instruction and output format are given first, followed by labeled video clips in cyclic order, with a blank boundary frame between consecutive clips. We use `top_p=1.0` and temperature 0.9 unless otherwise specified.

We set the requested summary length according to reference granularity: ActivityNet-derived references average about 3.4 sentences per video, so we ask for four sentences per video; News references average about 1.2 sentences, so we ask for two sentences per video. For frame-based inputs, we uniformly sample frames, using 16 frames per video for short-video settings and 24 frames per video for longer ActivityNet settings when supported. The default resolution is 448×448.

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 the aligned per-video summary.

3.3 Evaluation metrics

Let P be the number of input videos and $p \in \{1, \ldots, P\}$ the slot position. We denote the coverage score of the video at position p in instance i by $C_{i,p}$.

3.3.1 Coverage.

Coverage measures whether the generated summary preserves the key information in the human-written reference. Because reference granularity differs across datasets, we use dataset-adaptive estimators.

For ActivityNet-based domains, we use sentence-level semantic coverage. Given the reference sentences and the generated summary for the same video, GPT-5.1 OpenAI (2025) judges whether each reference sentence is covered, allowing paraphrases and minor wording differences. For News, whose references are much shorter and often concentrate key information in named entities, locations, organizations, and event phrases, we instead use reference-fragment coverage inspired by NEWSROOM Grusky et al. (2018), greedily matching the longest contiguous reference spans appearing in the generated summary. The two Coverage estimators are

$$C_{i,p} = \frac{N_{\text{cov}}^{(i,p)}}{N_{\text{ref}}^{(i,p)}}, \qquad C_{i,p}^{\text{frag}} = \frac{|\mathcal{M}_{i,p}|}{|r_{i,p}|}. \tag{1}$$

Here, $N_{\text{cov}}^{(i,p)}$ and $N_{\text{ref}}^{(i,p)}$ denote the number of covered and total reference sentences, respectively; $\mathcal{M}_{i,p}$ is the set of matched reference-token positions, and $|r_{i,p}|$ is the number of reference tokens.

3.3.2 Directional positional bias.

Directional Positional Bias (DPB) measures earlier-versus-later preference. We assign each position a linear weight

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

where $\alpha_1 = -1$ and $\alpha_P = 1$. For instance i:

$$\text{DPB}_i = \frac{\sum_{p=1}^{P} \alpha_p C_{i,p}}{\sum_{p=1}^{P} C_{i,p} + \epsilon}. \tag{3}$$

Model	Duration	Cooking-2	Cooking-4		Domestic-2	Domestic-4		Leisure-2	Leisure-4		News-2	News-4	
		DPB (%)	DPB (%)	MEG (%)	DPB (%)	DPB (%)	MEG (%)	DPB (%)	DPB (%)	MEG (%)	DPB (%)	DPB (%)	MEG (%)
Qwen3-VL-8B	Short	-3.33	-2.67	-1.34	8.13	-0.37	-2.57	6.01	0.67	-0.64	0.94	1.17	0.15
Qwen3-VL-8B	Long	–	–	–	-5.24	-2.90	-2.45	0.71	-2.37	-0.71	–	–	–
InternVL3.5-8B	Short	-3.34	-4.16	-3.34	-3.27	1.67	-4.50	-0.15	-2.07	-0.10	1.67	-0.99	-1.04
InternVL3.5-8B	Long	–	–	–	2.27	-0.94	-3.34	13.89	-1.28	-0.57	–	–	–
InternVL3.5-14B	Short	-5.93	-1.74	0.40	4.06	-2.51	-2.11	-3.04	0.07	-2.65	1.61	0.92	-1.49
InternVL3.5-14B	Long	–	–	–	-4.96	-0.98	-3.15	-8.35	-0.77	-1.41	–	–	–
Qwen3-VL-30B-A3B	Short	-5.68	-1.41	2.41	0.42	3.26	-2.63	9.36	-0.08	-0.76	2.11	0.09	-0.51
Qwen3-VL-30B-A3B	Long	–	–	–	5.01	-0.89	-3.68	-2.96	1.50	-1.23	–	–	–
MiniCPM-o-4.5	Short	-1.44	-1.56	-0.51	-2.69	1.56	1.92	3.06	1.82	0.21	-1.31	-0.50	-1.00
MiniCPM-o-4.5	Long	–	–	–	-4.48	1.35	-1.94	-12.05	-2.33	-1.96	–	–	–
GLM-4.1V-9B-Thinking	Short	-0.78	-2.06	-2.35	0.05	0.40	-1.22	4.87	-1.10	-2.82	-0.89	-0.72	-1.46
GLM-4.1V-9B-Thinking	Long	–	–	–	-0.51	1.08	0.79	-7.88	-5.70	-2.34	–	–	–
MiMo-VL-7B-RL	Short	14.69	-0.04	-0.95	-2.82	3.11	2.77	0.65	3.40	-4.95	0.38	0.36	-0.82
MiMo-VL-7B-RL	Long	–	–	–	-2.29	3.51	3.07	6.39	3.92	0.01	–	–	–
Gemini-3.1-Pro	Short	2.54	3.72	-0.73	17.88	2.48	0.47	-1.81	2.05	0.05	0.05	0.60	-0.18
Gemini-3.1-Pro	Long	–	–	–	-0.39	4.11	1.08	3.39	7.07	-0.43	–	–	–
GPT-5.4	Short	-1.79	-3.15	1.52	-14.97	1.86	-0.78	6.49	3.22	1.20	-0.29	-0.86	0.53
GPT-5.4	Long	–	–	–	1.83	0.72	0.48	4.83	1.16	-2.01	–	–	–

Table 1: Coverage-based Directional Positional Bias (DPB) and Middle-Edge Gap (MEG) across domains and models. Negative DPB values indicate primacy, i.e., higher coverage for earlier positions, while positive values indicate recency, i.e., higher coverage for later positions. MEG is defined only for four-video inputs; negative MEG values indicate middle-position weakness.

We aggregate over instances by averaging DPB_i across the N instances. Negative values indicate earlier-position preference, and positive values indicate later-position preference. Near-zero DPB does not imply position invariance, since non-monotonic effects, such as middle-position weakness, can cancel out in this signed measure. We therefore also use Middle-Edge Gap (MEG) to capture whether middle positions underperform edge positions in four-video inputs.

3.3.3 MIDDLE-EDGE GAP.

For four-video inputs ($P = 4$), we define MEG as:

$$\mathrm{MEG}_i = \frac{C_{i,2} + C_{i,3}}{2} - \frac{C_{i,1} + C_{i,4}}{2}. \tag{4}$$

We analogously report the mean of MEG_i over the N instances. Negative MEG indicates middle-position weakness, while positive MEG indicates a middle-position advantage.

4 EXPERIMENTAL RESULTS

4.1 MAIN RESULTS

Table 1 shows that positional effects do not reduce to a single universal primacy or recency pattern. Instead, the dominant pattern is heterogeneity across models and domains.

Middle-position weakness can be hidden by signed directional bias. DPB alone does not capture non-monotonic position effects. Several settings with near-zero DPB still exhibit negative Middle-Edge Gap (MEG). For instance, InternVL3.5-14B on Leisure-4 (Short) has DPB 0.07% but MEG -2.65%. Similarly, GLM-4.1V-9B-Thinking has DPB -1.10% but MEG -2.82%.

Pairwise and listwise settings reveal different behavior. Pairwise directional bias does not reliably predict four-video behavior. Several large $P{=}2$ effects shrink or reverse under $P{=}4$ inputs. For example, Qwen3-VL-8B changes from a strong recency pattern on Domestic-2 (Short, $+8.13\%$) to nearly zero DPB on Domestic-4 (Short, -0.37%). InternVL3.5-8B moves from strong recency on Leisure-2 (Long, $+13.89\%$) to mild primacy on Leisure-4 (Long, -1.28%). This suggests that listwise multi-video summarization introduces position interactions that are not visible in pairwise evaluation.

Video duration changes the shape of positional effects. This can occur because longer-duration settings use a larger sampled-frame budget, shifting the absolute input locations of later clips even when their relative slot indices are unchanged. In Domestic-4, several models show more negative MEG under longer inputs, such as Qwen3-VL-30B-A3B (-2.63% to -3.68%) and InternVL3.5-14B

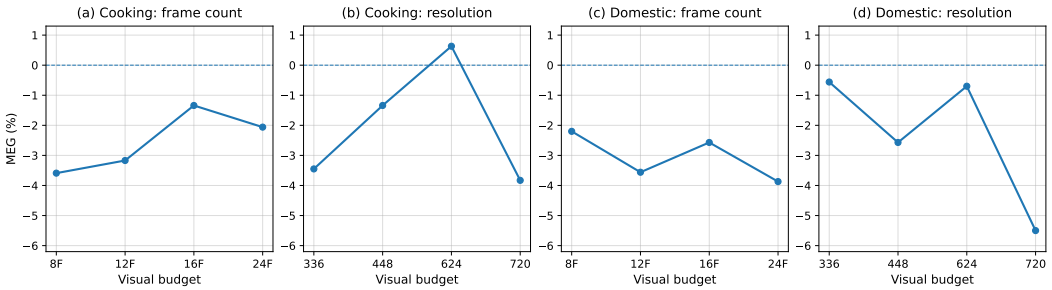


Figure 2: Visual-budget robustness check for Qwen3-VL-8B.

(-2.11% to -3.15%). In Leisure-4, negative MEG is common in both short and long settings, but some models change substantially: GPT-5.4 moves from positive MEG in the short setting ($+1.20\%$) to negative MEG in the long setting (-2.01%). These patterns suggest that duration and visual context length can modulate positional effects, rather than simply amplifying a fixed primacy or recency bias.

Case study. Among the proprietary models, we observe strong directional positional effects in both Gemini-3.1-Pro and GPT-5.4 on Domestic-2 (Short). GPT-5.4 has DPB -14.97% on Domestic-2 (Short), indicating substantially lower coverage for the second-position video. Manual inspection shows that the lower-scoring second-position summaries are often more generic, omitting concrete actions or object details. For example, summaries describing workers near a concrete or plaster machine were not judged as covering the reference phrase about men making plaster, and summaries mentioning a pit stop or loose tires were not judged as covering the specific action of changing tires.

This case also shows that small DPB values should be interpreted carefully. Near-zero DPB may reflect balanced coverage, but it may also occur when a model is uniformly generic and fails to cover all videos well. Stronger models may therefore exhibit larger measured positional bias because they preserve more details in favored positions, making uneven coverage easier to detect.

We also find that coverage differences can be accompanied by uneven output allocation. Table 2 shows Qwen3-VL-8B on Domestic-4 (Short). The middle slots have comparable word counts to the first slot but lower coverage, while the final slot is longer and more often exceeds the requested number of sentences.

4.2 Visual-budget robustness.

We further examine whether the observed positional effect is an artifact of a particular visual input configuration. This analysis uses Qwen3-VL-8B on the $P{=}4$ short-video setting. We vary the number of sampled frames while fixing the resolution at $448{\times}448$, and vary the resolution while fixing the frame count at 16. We report MEG (%), where negative values indicate that middle positions underperform edge positions.

Figure 2 shows that the negative MEG pattern is largely preserved under visual-budget changes. In *Domestic*, MEG remains negative under all tested settings.

We also run the same visual-budget ablation on MiniCPM-o-4.5, for which frame count mainly affects overall Coverage. We report the full MiniCPM-o-4.5 ablation in Appendix C.4.

4.3 Boundary-format robustness

Prior multi-video benchmarks show that MLLMs can struggle to filter irrelevant information across videos in multi-video inputs Peng et al. (2025). This makes input boundary design a relevant protocol factor: visual separators may help models segment videos, while text-only labels may change how strongly the model separates one video from another. We therefore compare the default black-frame separator with a text-label-only variant.

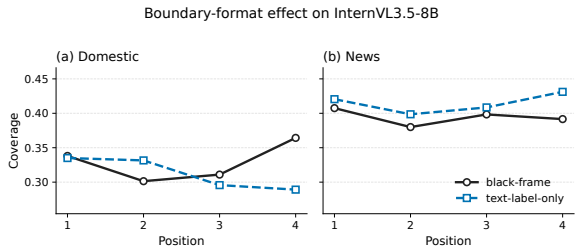


Figure 3: Boundary-format comparison for InternVL3.5-8B.

Position	Avg. words	Over-rate	Coverage (%)
Pos. 1	66.92	2.78%	43.88
Pos. 2	66.60	1.85%	41.18
Pos. 3	67.09	6.48%	40.47
Pos. 4	69.68	8.33%	42.90

Table 2: **Output allocation example for Qwen3-VL-8B on Domestic-4 (Short).** Over-rate denotes the proportion of outputs exceeding the requested number of sentences.

Setting	Held-out gap (%)	95% CI
Cooking Short	10.21	[7.11, 13.54]
Domestic Short	7.48	[4.87, 10.17]
Leisure Short	7.28	[3.97, 10.70]
News Short	8.14	[7.04, 9.30]
Domestic Long	11.03	[7.70, 14.38]
Leisure Long	4.06	[1.59, 6.57]

Table 3: Leave-one-repeat-out position contrasts for MiniCPM-o-4.5 at $P{=}4$.

Scope	MEG (%)	95% CI	p
All $P{=}4$	-0.944	$[-1.358, -0.520]$	0.00005
ActivityNet	-1.004	$[-1.496, -0.494]$	0.00020
News	-0.646	$[-1.010, -0.265]$	0.00205

Table 4: Suite-level middle-edge effects using original video groups as the resampling unit.

Overall, changing the boundary format affects the position-wise coverage curve, but does not provide a consistent improvement. For InternVL3.5-8B, the text-label-only format changes the curve more visibly: *Cooking* and *Domestic* show less negative MEG, but this does not translate into uniformly better summary quality. In *Domestic*, position 2 improves while position 4 drops substantially, suggesting that text-only boundaries flatten the curve rather than consistently improving coverage. For Qwen3-VL-8B, removing black frames does not remove the middle-position weakness.

As a slot-boundary sanity check, we also compare each generated slot summary with its own reference and with the other references in the same group using TF-IDF content-token similarity. A slot is marked as potential contamination only when it is substantially more similar to another video's reference than to its own. This proxy is not a verified leakage label, but it can reveal large-scale slot-boundary failures. We do not observe a systematic increase in potential contamination under text-label-only formatting.

4.4 Reliability and Reproducibility

We next examine whether the positional effects observed above remain statistically detectable and reproducible across independent generations. Three repeated MiniCPM-o-4.5 runs collected under the same inference protocol provide a direct test of cross-run stability. For each target video, we estimate its relative position profile from two runs and evaluate the corresponding position contrast on the held-out third run, rotating the held-out run across all three choices.

Across the six $P{=}4$ settings, relative position preferences estimated from two runs remain predictive for the held-out generation (Table 3). The held-out contrasts range from 4.06% to 11.03% in Coverage, indicating that the observed position-conditioned variation contains a reproducible component across independent generations. The remaining differences across settings may arise from the interaction of surrounding-video context, target binding, and autoregressive output allocation.

At the benchmark level, the same conclusion is supported by group-level statistical analysis. Treating the nine evaluated models as a fixed benchmark suite and using the original video group as the resampling unit, the pooled $P{=}4$ MEG is -0.944% (95% CI $[-1.358, -0.520]$; group-level sign-flip $p = 0.00005$). The effect is also detected independently with both evaluation families. ActivityNet-derived domains, evaluated with the LLM judge, yield a pooled MEG of -1.004%, while News, evaluated using reference-fragment token matching, yields -0.646% (Table 4). This agreement across

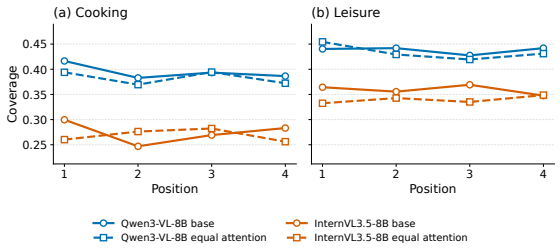


Figure 4: Effect of equal-attention prompting on position-wise coverage for Qwen3-VL-8B and InternVL3.5-8B.

Target pos.	ΔCoverage	ΔTarget hit	ΔMargin confusion
Pos. 1	−0.54	−0.23	−0.93
Pos. 2	+4.18	+14.12	−11.81

Table 5: Effect of changing single-target labels from zero-indexed to one-indexed form on InternVL3.5-14B.

distinct evaluation procedures strengthens the evidence for an average middle-position disadvantage in four-video summarization.

Validation of the cyclic ordering design. The stored full-permutation outputs further allow us to evaluate the four-order cyclic protocol against the complete 24-permutation set. Under the same four-order inference budget, balanced cyclic subsets yield an average MEG estimation error of 1.21%, compared with 2.20% for four randomly sampled permutations, corresponding to an approximately 45% reduction. This advantage is consistent across the two models with full-permutation records: 1.22% versus 2.12% for MiniCPM-o-4.5 and 1.21% versus 2.28% for Qwen3-VL-8B.

5 Prompt-level mitigation and diagnostics

We explore lightweight prompt-level interventions and diagnostics that do not require model retraining or architectural access. We first test an explicit balanced-attention instruction, then evaluate a single-target generation protocol, and finally examine prompt placement as a target-binding diagnostic.

5.1 Balanced-attention instruction.

We first add a direct instruction asking the model to allocate equal attention to each video. Figure 4 shows that this intervention does not provide reliable mitigation. For Qwen3-VL-8B, the equal-attention prompt does not produce a stable balancing effect on Coverage across positions. For InternVL3.5-8B, the Cooking curve becomes flatter, but this flattening partly comes from lowering edge-position coverage rather than uniformly improving all positions. Thus, the prompt can reshape the position-wise coverage curve, but does not consistently improve summary quality or remove positional imbalance.

5.2 Single-target prompting

We also test a single-target generation protocol as an output-side diagnostic. The model receives the same multi-video input, but each call requests only one specified target-video summary instead of all summaries in one response. This removes previously generated summaries from the autoregressive prefix and tests whether position-dependent errors arise from joint generation effects, such as uneven detail allocation, or cross-summary contamination. If the bias persists, it is more likely due to the multi-video input itself; if it weakens, joint multi-summary generation may contribute to the effect.

Target-addressing concern. Single-target prompting introduces an additional target-addressing ambiguity that is less salient in joint per-video generation. Unlike joint generation, where the model outputs descriptions for all videos together, single-target prompting requires the model to resolve one specified label. We observe that models may confuse this label with ordinal position: even when the prompt explicitly states which frames correspond to each label, `[video1]` can still be treated as the first or currently queried video rather than the second zero-indexed video.

To quantify this label-resolution issue, we compute TF-IDF cosine similarity between each generated summary and the four references in the same group. A target hit occurs when the true target reference

Scale	ΔWeak Cov.	ΔWeak Sent.	Weak pos. imp.	ΔDensity
P=2	−4.09	+0.16	0/6	−0.121
P=4	+2.87	+1.26	4/6	−0.224

Table 6: Effect of single-target prompting relative to joint generation for InternVL3.5-14B. Weak positions are identified from the joint-generation setting. ΔDensity is the change in coverage per 100 words.

Protocol	Cand. sent.	Verified leak	Cand. prec.	Slot rate	Sent. rate
Prompt-first	263	52	19.77%	1.93%	0.68%
Video-first	446	96	21.52%	3.56%	0.93%

Table 7: Manual verification of candidate cross-video leakage. Cand. sent. denotes mined candidate sentences; Verified leak denotes manually confirmed leakage sentences; Cand. prec. is Verified leak / Cand. sent.

has the highest similarity; a margin confusion occurs when another reference exceeds the true target by over 0.05. Table 5 shows that one-indexed labels mainly improve target addressing at position 2, where zero-indexing is most ambiguous. We therefore compare zero-indexed and one-indexed labels on InternVL3.5-14B to test whether this addressing convention contributes to the single-target effect.

Effect analysis. Table 6 shows a scale-dependent effect. For P=4, single-target prompting improves the weak-position coverage by 2.87% and improves the weakest joint-generation position in 4 of 6 datasets. The gain is accompanied by a much larger increase in sentence count, especially for positions that were compressed in the joint output.

In contrast, for P=2, the weak-position coverage decreases despite an increase in generated sentences. Overall, this protocol is not a general improvement, but a possible inference-time workaround when the number of input videos grows and joint generation begins to compress some summary slots.

5.3 Prompt placement and leakage audit

We next examine whether prompt placement affects target binding. In the main setting, the task instruction appears before the videos, while a video-first variant places the same instruction after the videos. We first inspect the easier P=2 setting, where inputs are shorter and model performance is generally stronger. For InternVL3.5-14B, video-first produces slightly longer ActivityNet outputs on average (3.817 vs. 3.667 sentences), but reduces P=2 coverage by 4.29%. This suggests that prompt placement affects target binding, not merely instruction compliance.

To further test this target-binding concern, we conduct a targeted leakage audit. A full target-alignment audit would require comparing thousands of generated summaries with their corresponding videos. Instead, we use a reference-based candidate miner to reduce the review space. The miner is designed around the observation that cross-video leakage often appears as a concrete action or object detail from another video, even when clips have similar scenes. For each generated sentence, it compares concrete action/object anchor words against the reference for the claimed target video and the references of the other videos in the same group. A sentence is selected as a candidate when its concrete details are weakly supported by the claimed target but better supported by another video.

Table 7 shows that video-first prompting produces more mined candidates and more manually verified leakage sentences than prompt-first prompting. At the position level, candidate precision is higher for positions 2 and 3 than for positions 1 and 4 under both prompt placements. We observe two recurring failure modes: full target-addressing failures, where the summary is assigned to the wrong video, and local detail intrusions, where the summary mostly describes the target video but borrows a concrete action or object from another video. Since one target-addressing error can affect multiple generated sentences, sentence-level counts may overstate the number of independent leakage events.

6 Conclusion

We present a systematic study of positional bias in multi-video summarization for MLLMs. Our results show that current models are not reliably position-invariant: summary quality varies with input order across domains, model families, and input settings. We further show that signed directional bias alone is insufficient, since low DPB can still hide middle-position weakness under MEG. Robustness and protocol analyses suggest that these effects are not reducible to a single artifact, while lightweight prompt-level interventions only partially reshape them. Our benchmark, metrics, and diagnostics provide a foundation for studying and improving order robustness in multi-video summarization.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing of the manuscript and with debugging and refining code that was initially written by the authors for the experimental pipeline. All AI-assisted text, code suggestions, and analysis outputs were manually inspected and revised by the authors before being incorporated into the paper or experiments.

REFERENCES

Sacha Altay, Emma Hoes, and Magdalena Wojcieszak. Following news on social media boosts knowledge, belief accuracy and trust. *Nature Human Behaviour*, 9:1833–1842, 2025. doi: 10.1038/s41562-025-02205-6.

Purui Bai, Tao Wu, Jiayang Sun, Xinyue Liu, Huaibo Huang, and Ran He. Mvpbench: A multi-video perception evaluation benchmark for multi-modal video understanding, 2026. URL https://arxiv.org/abs/2603.22756.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15 (3), March 2024. ISSN 2157-6904. doi: 10.1145/3641289. URL https://doi.org/10.1145/3641289.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

Google DeepMind. Gemini 3.1 pro model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6019–6024, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1620. URL https://aclanthology.org/D19-1620/.

Max Grusky, Mor Naaman, and Yoav Artzi. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 708–719, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1065. URL https://aclanthology.org/N18-1065/.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, 2015. doi: 10.1109/CVPR.2015.7298698.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems, 2024. URL https://arxiv.org/abs/2305.08845.

Woojun Jung and Junyeong Kim. QEVA: A reference-free evaluation metric for narrative video summarization with multimodal question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24632–24642, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1340. URL https://aclanthology.org/2025.findings-emnlp.1340/.

Kajal Kansal, Nikita Kansal, Sreevaatsav Bavana, Bodla Krishna Vamshi, and Nidhi Goyal. 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, pp. 65–73, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702778. doi: 10.1145/3607540.3617139. URL https://doi.org/10.1145/3607540.3617139.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Lin Liang and Shasha Wang. Exploring learning engagement among spanish undergraduates in asia: the chain mediating roles of flow experience and self-determined motivation in short video platform use. *Frontiers in Psychology*, 17, 2026. doi: 10.3389/fpsyg.2026.1709572.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

OpenAI. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf, 2025.

OpenAI. GPT-5.4 Thinking System Card. https://openai.com/index/gpt-5-4-thinking-system-card/, March 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Zekun Moore Wang, Zili Wang, Ge Zhang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, Houyi Li, Wei Ji, Pengfei Wan, Wenhao Huang, Zhaoxiang Zhang, and Jiaheng Liu. MVU-eval: Towards multi-video understanding evaluation for multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=UZD5CQV6f9.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8029–8036, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.633. URL https://aclanthology.org/2021.emnlp-main.633/.

Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20643–20664, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1124. URL https://aclanthology.org/2025.findings-emnlp.1124/.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *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*, pp. 292–314, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. URL https://aclanthology.org/2025.ijcnlp-long.18/.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

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

David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. On positional bias of faithfulness for long-form summarization. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 8791–8810, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.442. URL https://aclanthology.org/2025.naacl-long.442/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. Incorporating background knowledge into video description generation. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3992–4001, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1433. URL https://aclanthology.org/D18-1433/.

Hou Xia, Zheren Fu, Fangcan Ling, Jiajun Li, Yi Tu, Zhendong Mao, and Yongdong Zhang. Videolevelgauge: Investigating contextual positional bias in large video language models, 2025. URL https://arxiv.org/abs/2508.19650.

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 60755–60775. Curran Associates, Inc., 2024. doi: 10.52202/079017-1943. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/6ffdbbe354893979367f93e2121e37dd-Paper-Conference.pdf.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

A PROMPTS AND LLM-AS-A-JUDGE

A.1 PER-VIDEO SUMMARIZATION PROMPTS

We use a fixed per-video output structure to support reliable slot extraction and position-wise analysis. In the main experiments, the task prompt is placed before the visual inputs. For ActivityNet-based domains, we request four sentences per video. For the News domain, whose references are much shorter, we request two sentences per news video. The required output markers are zero-indexed and extend to the number of videos in the input, e.g., `[video0]` through `[video3]` for four-video inputs.

Main black-frame prompt.

```
Different videos in the input are separated by black frames.
Summarize each video in 4 sentences.
Use exactly this format, including the square brackets:
[video0] summary
[video1] summary
...
[video{P-1}] summary
```

Input layout:

```
Video 0: {N} uniformly sampled frames
{frames of video 0}
<black separator>
Video 1: {N} uniformly sampled frames
{frames of video 1}
<black separator>
...
Video {P-1}: {N} uniformly sampled frames
{frames of video {P-1}}
```

Text-label-only prompt.

```
Different videos in the input are separated by text labels only.
Summarize each video in 4 sentences.
Use exactly this format, including the square brackets:
[video0] summary
[video1] summary
...
[video{P-1}] summary
```

Input layout:

```
Video 0: {N} uniformly sampled frames
{frames of video 0}
— Text separator: end of Video 0; next is Video 1. —
Video 1: {N} uniformly sampled frames
{frames of video 1}
— Text separator: end of Video 1; next is Video 2. —
...
Video {P-1}: {N} uniformly sampled frames
{frames of video {P-1}}
```

For News, the second line is replaced by `Summarize each news video in 2 sentences.`

A.2 LLM-AS-A-JUDGE PROMPT FOR SENTENCE-LEVEL COVERAGE

We evaluate ActivityNet Coverage using an LLM-as-a-Judge protocol with `gpt-5.1-2025-11-13`. Given a reference summary split into sentences and a model-generated summary for the same video, the judge returns a binary covered/not-covered decision for each reference sentence. We use the following prompt with temperature 0.

```
You are evaluating whether a model-generated video summary covers a human reference
summary.
For EACH reference sentence, decide whether it is clearly covered by the model-generated
summary.

Scoring rule:
- covered = 1 if the model summary explicitly or implicitly describes the same fact as
the reference sentence.
- covered = 0 if the model summary omits the key information, contradicts the reference
sentence, or is too vague/general to support the specific reference sentence.

Rules:
- Different wording is acceptable when the meaning matches.
- Do not require exact phrasing.
- If only part of a reference sentence is covered and an important fact is missing,
score 0.
- For covered=0, explain briefly what is missing or contradicted.
- For covered=1, reason must be an empty string.

Reference sentences:
{numbered_refs}

Model-generated summary:
{model_summary}

Return ONLY valid JSON. Do not include markdown.
Required JSON schema:
{
"sentence_scores": [
{"index": 1, "covered": 1, "reason": ""},
{"index": 2, "covered": 0, "reason": "The model summary does not mention ..."}
]
}
```

A.3 Human agreement and judge robustness

To assess the reliability of the sentence-level coverage judge, we recruited three undergraduate annotators and compensated them at a rate of USD 8 per hour. On the same 245 summaries with valid judge outputs, we compute pairwise human–human agreement and the agreement between the mean human score and the LLM judge. Table 8 reports Pearson and Spearman correlations.

Comparison	n	Pearson	Spearman
Annotator 1 vs. Annotator 2	245	0.7650	0.7519
Annotator 1 vs. Annotator 3	245	0.7007	0.6969
Annotator 2 vs. Annotator 3	245	0.7609	0.7637
Human mean vs. LLM	245	0.7582	0.7419

Table 8: Human agreement and LLM-judge robustness on the same 245 summaries.

The LLM judge shows strong agreement with the mean human score (Pearson $r = 0.7582$, Spearman $\rho = 0.7419$), comparable to the observed pairwise human–human correlations.

Model	Marker fail.	Fallback
Qwen3-VL-8B	1	1
InternVL3.5-8B	5	1
InternVL3.5-14B	8	0
Qwen3-VL-30B-A3B	1	0
MiMo-VL-7B-RL	14	3

Table 9: Marker parsing failures in the baseline records.

Setting	P=2	P=4
Cooking	3	22
Domestic-S	6	24
Domestic-L	7	22
Leisure-S	10	28
Leisure-L	8	36

Table 10: MiMo-VL-7B-RL complete output failures.

B 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 Computational budget.

All local experiments were conducted on NVIDIA RTX PRO 6000 GPUs with 96 GB of memory. Qwen3-VL-30B-A3B and InternVL3.5-14B were run with three GPUs, while the other locally evaluated models were run with a single GPU. Across local inference and automatic evaluation runs, the total computational budget was approximately 1,100 GPU-hours.

C.2 Parsing and output-failure diagnostics

We also inspect whether positional effects are confounded by parsing or complete output failures. Most models follow the requested `[videoX]` format. We distinguish two failure modes. A *marker failure* occurs when the model attempts the fixed marker format but omits one or more required video slots or uses malformed slot markers; such cases are marked as failures for the affected slots. A *fallback* case occurs only when the model does not follow the fixed marker format at all, in which case we use BERTScore sentence assignment to recover a best-effort alignment. Table 9 reports only failure counts.

Complete output failures are more concentrated. Table 10 shows the MiMo-VL-7B-RL settings where the model produced placeholder-only outputs, such as literal `[video0] summary`-style text, rather than substantive descriptions of the videos. These outputs are unusable for coverage evaluation and are assigned zero Coverage in the baseline aggregation. These failures occur mainly in ActivityNet settings and become more frequent in four-video inputs.

C.3 Coverage curves for all baseline models

Figures 5 and 6 report the full baseline Coverage–position curves. We split the models into two groups for readability; both figures use the same ordering of domains, durations, and input scales.

Some Gemini-3.1-Pro settings show relatively low Coverage despite producing fluent summaries. This reflects the intended scope of our metric rather than a global judgment of summary quality. Coverage is reference-grounded: since summaries are ultimately intended for human readers, we assume that human-written references identify the salient facts and events that readers expect a summary to preserve. Our inspection suggests that Gemini-3.1-Pro often produces shorter and more abstractive summaries, which may contain fewer explicit reference-grounded details under this metric. This limitation should be considered when interpreting the positional analysis, because we compare Coverage across input slots under the same scoring rule. If reference-grounded information preservation varies by slot, the model still exhibits order-dependent summarization behavior.

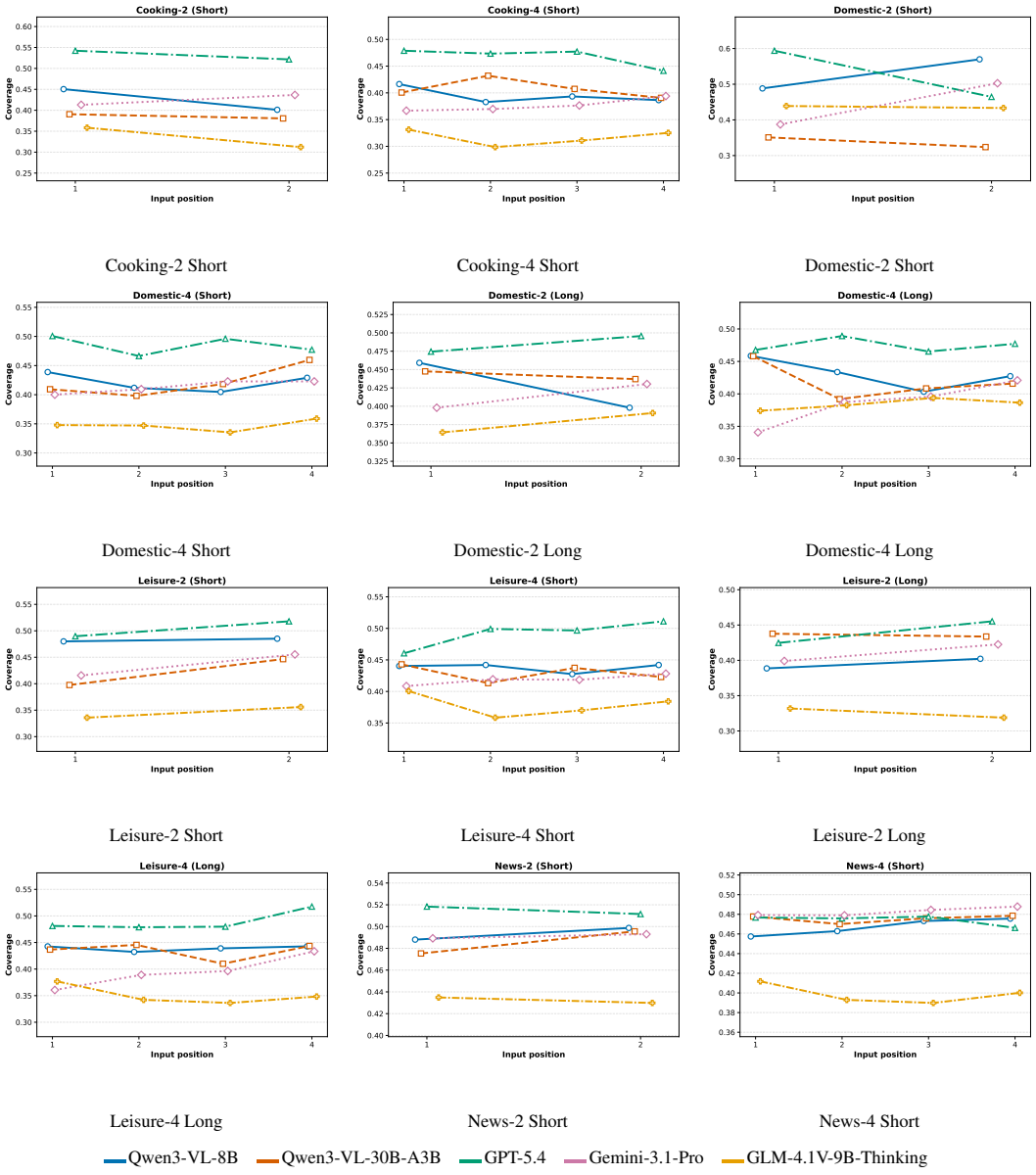


Figure 5: Baseline Coverage–position curves for model group A. Each panel corresponds to one domain, duration, and input-size setting.

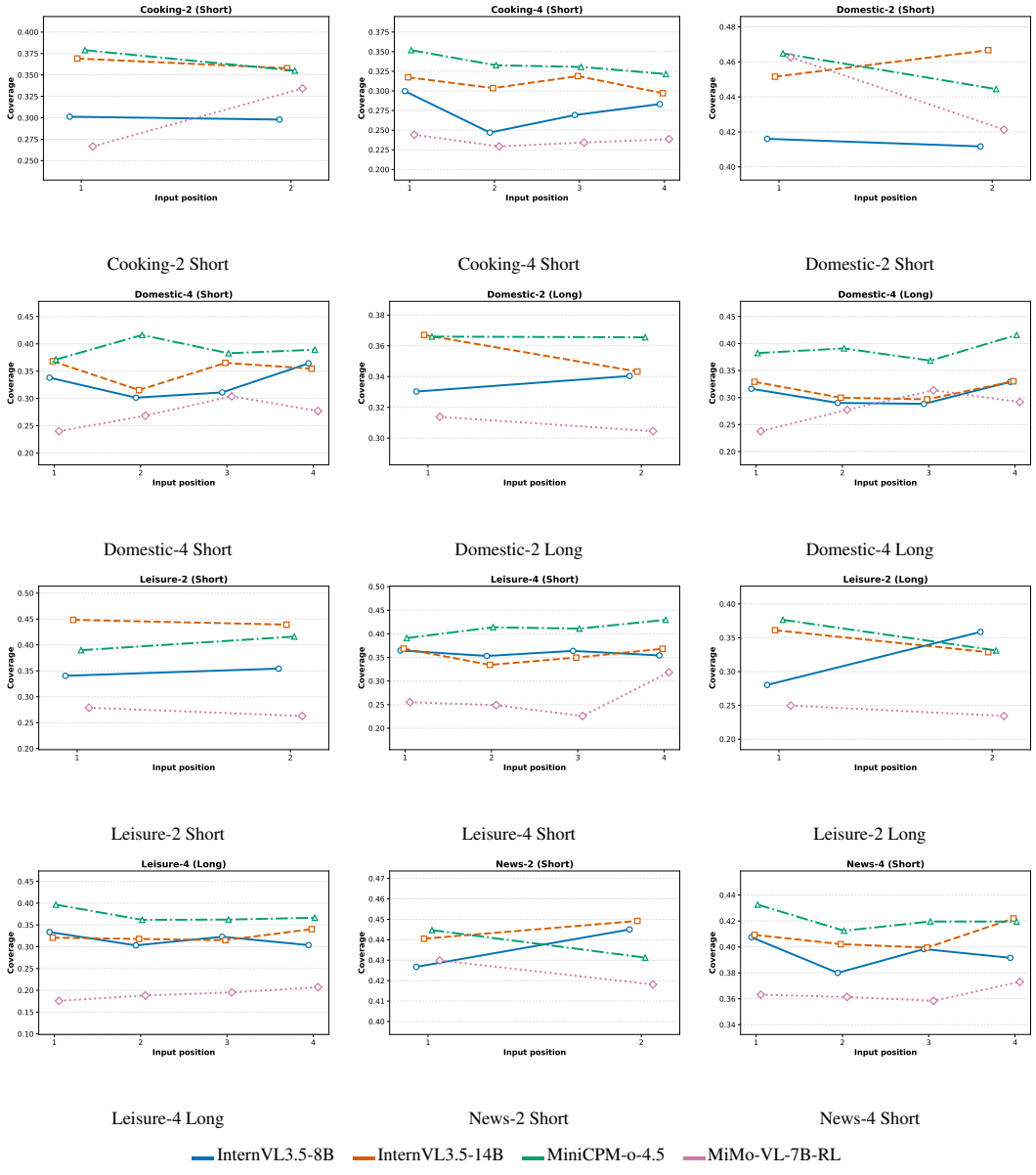


Figure 6: Baseline Coverage–position curves for model group B. The panel order matches Figure 5.

(a) Qwen3-VL-8B

Domain	Metric	Base	8F	12F	24F	336R	624R	720R
Cooking	Cov.	39.48	32.07	36.73	38.18	35.62	41.03	38.90
	DPB	-1.68	2.55	-0.19	-0.27	0.51	1.38	1.80
	MEG	-1.34	-3.59	-3.17	-2.06	-3.45	0.63	-3.83
Domestic	Cov.	42.11	37.32	40.47	42.59	39.01	41.09	42.64
	DPB	-0.72	-2.70	0.69	0.07	-1.17	3.57	-0.81
	MEG	-2.57	-2.20	-3.56	-3.87	-0.56	-0.70	-5.50
Leisure	Cov.	43.79	38.08	39.81	43.68	42.33	44.23	43.71
	DPB	-0.18	2.22	-2.23	0.37	0.48	0.83	0.53
	MEG	-0.64	-1.13	-1.49	-0.19	-3.12	-1.89	-0.10
News	Cov.	46.73	37.14	44.78	47.04	45.92	46.27	45.87
	DPB	1.16	0.07	0.95	0.85	0.51	0.21	0.37
	MEG	0.15	1.27	0.49	-0.24	0.34	-1.00	1.30

(b) MiniCPM-o-4.5

Domain	Metric	Base	8F	12F	24F	336R	624R	720R
Cooking	Cov.	34.23	30.01	34.48	38.04	32.70	34.93	35.49
	DPB	-0.58	0.49	1.00	-0.39	-0.30	0.39	-0.42
	MEG	-2.03	-2.42	-1.79	0.40	-0.65	-3.06	1.02
Domestic	Cov.	39.35	34.38	34.96	40.85	41.93	41.84	40.83
	DPB	0.79	0.90	-2.27	0.47	-0.35	-3.42	1.61
	MEG	0.86	-1.45	0.52	-1.35	-0.58	-1.48	3.31
Leisure	Cov.	40.90	35.00	39.54	42.70	40.42	40.89	40.29
	DPB	-0.32	-2.13	-1.84	-1.10	0.64	-0.43	-0.00
	MEG	-0.85	0.81	-0.96	-1.05	-1.34	-2.36	0.01
News	Cov.	42.05	31.20	39.01	42.51	41.54	41.43	42.05
	DPB	-0.45	-2.45	-1.17	0.19	-0.60	-0.00	-0.58
	MEG	-1.19	-0.78	0.60	-1.18	-0.16	-1.04	-0.10

Table 11: Visual-budget ablation summary for Qwen3-VL-8B and MiniCPM-o-4.5 on short $P{=}4$ settings. All values are percentages. Base denotes 16 frames at $448{\times}448$. Frame-count variants fix resolution at $448{\times}448$; resolution variants fix frame count at 16.

Budget	Requested sent.	All outputs	Exact-budget
Short	2	-2.29	-0.63
Base	4	-2.45	-2.60
Long	6	-4.22	-3.70

Table 12: MEG for the Domestic-long setting under different requested summary budgets.

Domain	Dur.	Weakest position	Weak gain	Nonweak gain	Diff.
Cooking	S	Position 4	$+5.29$	-0.59	$+5.88$
Domestic	L	Position 3	-0.07	$+0.15$	-0.22
Domestic	S	Position 2	$+3.66$	$+0.23$	$+3.43$
News	S	Position 3	$+4.15$	$+2.18$	$+1.97$
Leisure	L	Position 3	-0.90	$+0.99$	-1.89
Leisure	S	Position 2	$+5.07$	$+4.36$	$+0.71$

Table 13: Weak-position gain for InternVL3.5-14B in $P{=}4$ settings. Gains are single-target minus joint generation (%). Diff. is weak-position gain minus the average gain of the remaining positions.

Domain	Dur.	P	Joint	Single	Gain (%)
Cooking	S	2	36.35	31.13	-5.22
Cooking	S	4	30.92	31.81	$+0.88$
Domestic	L	2	35.51	31.93	-3.59
Domestic	L	4	31.39	31.48	$+0.09$
Domestic	S	2	45.91	42.89	-3.02
Domestic	S	4	35.08	36.17	$+1.09$
News	S	2	44.49	43.84	-0.65
News	S	4	40.82	43.50	$+2.68$
Leisure	L	2	34.49	30.74	-3.75
Leisure	L	4	32.35	32.87	$+0.52$
Leisure	S	2	44.39	38.75	-5.63
Leisure	S	4	35.52	40.06	$+4.54$

Table 14: Dataset-level mean Coverage for InternVL3.5-14B under joint generation and single-target prompting. Gain is single-target minus joint generation (%).

C.4 Visual-budget ablation tables

The main text reports a visual-budget robustness check for Qwen3-VL-8B. Here we provide the full visual-budget summaries for Qwen3-VL-8B and MiniCPM-o-4.5, including Coverage, DPB, and MEG. For Qwen3-VL-8B, negative MEG is largely preserved across frame-count and resolution variants, especially in *Domestic*. MiniCPM-o-4.5 shows a different pattern: frame count mainly changes overall Coverage, with 8 frames lowering Coverage and 24 frames improving Coverage across domains, while MEG remains mixed. These results support the main-text conclusion that larger visual budgets do not necessarily remove positional imbalance.

C.5 Requested summary budget and middle-position weakness

We next test whether middle-position weakness is driven by the summary-length constraint in the prompt. A tight requested budget may force the model to omit details unevenly across positions. We therefore vary the requested per-video budget with three settings: short, base, and long, corresponding to 1/2, 2/4, and 3/6 requested sentences for News/ActivityNet-based domains, respectively. For each four-video input, we compute MEG and also report an exact-budget subset whose generated sentence counts match the requested numbers.

Table 12 illustrates the pattern on Domestic-long. Under the short budget, the gap becomes much smaller in the exact-budget subset, suggesting that a tight two-sentence budget compresses all

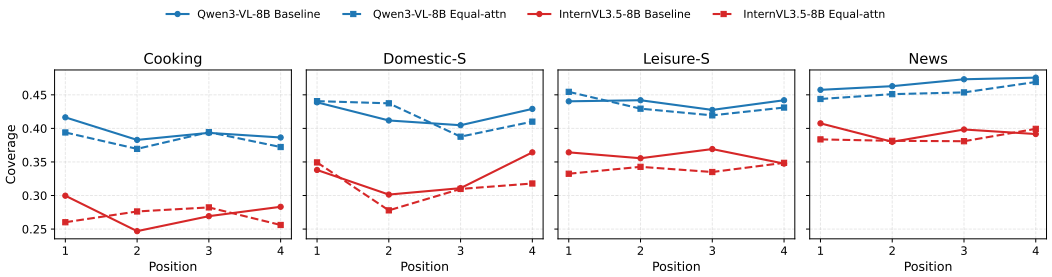


Figure 7: **Equal-attention prompt Coverage curves across all short four-video domains.** Solid lines are baseline prompt-first black-frame runs; dashed lines add the explicit equal-attention instruction.

positions and can mask position-specific differences. With larger budgets, the gap remains negative even after sentence-count matching, and is largest in the long setting. Thus, giving the model more output space does not necessarily remove middle-position weakness; it can instead make uneven detail allocation more visible.

More broadly, increasing the requested budget does not benefit all positions equally. From short to long, middle positions receive smaller coverage gains than edge positions in four of six $P = 4$ datasets. This suggests that the issue is not only whether the model writes enough sentences, but how added generation space is allocated across input positions.

C.6 InternVL3.5-14B single-target results across all datasets

Tables 14 and 13 provide the dataset-level results behind the single-target analysis in Section 5.2. The tables separate mean Coverage changes from weak-position changes. Single-target prompting lowers mean Coverage in all six P=2 settings, but improves mean Coverage in all six P=4 settings. The weak-position analysis shows the same scale dependence: none of the P=2 weak positions improves, while 4 of 6 P=4 weak positions improve.

C.7 Equal-attention coverage curves

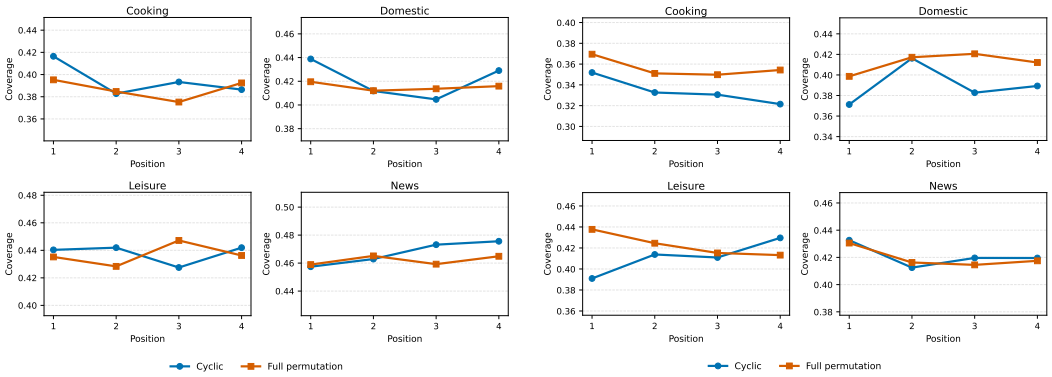
Figure 7 reports the equal-attention prompt intervention on all four short four-video domains for Qwen3-VL-8B and InternVL3.5-8B. These curves complement the two-domain visualization in the main text and show that the prompt intervention is model- and domain-dependent rather than a uniform fix for positional imbalance.

C.8 Cyclic versus full-permutation ordering

Our main experiments use four balanced cyclic orderings, which place every video at each absolute position exactly once. For four-video inputs, this reduces the ordering space from $4! = 24$ permutations to four evaluations per group while preserving complete target–position balance. We further use stored full-permutation outputs for Qwen3-VL-8B and MiniCPM-o-4.5 to examine how the estimated position curves change when a broader range of surrounding video contexts is averaged.

The full-permutation results show that positional effects are shaped jointly by absolute position and surrounding-video configuration. For Qwen3-VL-8B, MEG remains negative for Cooking under both designs, Domestic remains negative with a smaller magnitude under the full-permutation setting, and News stays close to zero; Leisure changes from a weak negative to a weak positive MEG. MiniCPM-o-4.5 shows a similar domain-dependent pattern: Cooking and News remain negative, Domestic remains positive, and Leisure changes from a weak positive to a negative MEG. These results indicate that cyclic ordering captures a useful low-budget view of positional behavior while full-permutation evaluation reveals additional position–context interactions.

The complete permutation sets also provide a direct validation of the four-order design. Using the 24-order average as the reference, balanced four-order cyclic subsets produce substantially smaller



(a) Qwen3-VL-8B (b) MiniCPM-o-4.5

Figure 8: **Cyclic versus full-permutation Coverage curves for Qwen3-VL-8B and MiniCPM-o-4.5 on the short $P{=}4$ settings.**

MEG estimation error than four randomly sampled permutations under the same inference budget. Averaged across the four short-video domains and the two models, the error is approximately 1.21% for the four-order cyclic subset and 2.20% for four randomly sampled permutations, a reduction of about 45%. Increasing the balanced budget to eight and twelve orders further reduces the error to approximately 0.77% and 0.54%, respectively. This supports balanced cyclic ordering as an efficient approximation to full-permutation evaluation under a limited inference budget.

D EXAMPLES

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 9: **Example 1.** The 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 at Position 1.

Prompt-placement leakage example.

Claimed target slot: Position 4, "Drinking coffee"



Source slot whose details leak in: Position 1, "Having an ice cream"



Target reference. A big green garbage can is pushed up against a brick wall and a reindeer jumps out drinking from a coffee cup. The reindeer continues to drink out of the coffee cup and eventually falls out of over the garbage can.

Source reference. A little boy is filling an ice cream cone with ice cream. He licks the ice cream cone. He lifts the ice cream cone up to show the camera.

Code-mined anchors specific to other videos. cone; cream; ice; ice cream cone; enjoy cream; ice cream; capture with cream.

Original model summary for the claimed target slot. An animated deer climbs out of a green dumpster, holding an ice cream cone with excitement. The background features a brick wall and a street scene, adding context to the urban setting. The deer attempts to enjoy its ice cream, even managing to catch some of it in its mouth playfully. The scene is whimsical, capturing the deer's joyful interaction with the ice cream.

Human label. Confirmed cross-video leakage: the claimed target video shows a deer/reindeer drinking from a cup, while the highlighted ice-cream anchors are supported by another video in the same input group.

Figure 10: **Manually verified prompt-placement leakage case.** The automatic miner selected the highlighted concrete anchors because they are unsupported by the claimed target reference but supported by another video in the same input group. The human audit labels this case as confirmed cross-video leakage.