Does Inference-Time Reasoning Really Improve Video Understanding?

Zhuohao Yu[1] **Yu Zhou**[2] **Zhecan Wang**[2]
Rui Sun[2] **Kai-Wei Chang**[2] **Nanyun Peng**[2]

[1]Carnegie Mellon University [2]University of California, Los Angeles

Abstract

Video understanding models achieve impressive accuracy on recent benchmarks, but do these benchmarks truly measure genuine grounded understanding? We systematically investigate this by controlling frame sampling across VideoMME, Video-MMMU, and LVBench, with matched thinking and no-thinking comparisons at fixed sampled frames. Our analysis identifies shortcuts in 33.33%, 41.00%, and 12.46% of questions, respectively, through world knowledge, language priors, or frame-level cues. Requiring both correct answers and temporal localization exposes a gap between answering questions and locating relevant evidence. To study this distinction, we introduce TGVMME, a temporally grounded benchmark with 1,500 questions requiring both answer correctness and temporal localization. Inference-time reasoning produces small or mixed effects under sparse input, while denser input permits larger, model-dependent gains. On existing benchmarks, reasoning gains are larger on detected shortcut questions in aggregate: +3.99 points compared with -2.84 points on the remaining questions, a +6.83-point contrast with a paired 95% confidence interval of [+5.68, +7.98]. These groups are identified from successful thinking responses and held fixed across the comparisons. For Qwen3-VL on TGVMME, increasing sampling from 0.2 to 1 frame per second yields +19.84 answer-accuracy points, compared with +1.44 points from enabling reasoning at the same frame rate. Our work demonstrates the value of grounded evaluation for assessing video understanding and provides a shortcut taxonomy, benchmark, and controlled analysis of visual input and inference-time reasoning.

1 Introduction

Video-language models achieve impressive 70%+ accuracy on benchmarks like VideoMME [13] and Video-MMMU [22], but *do these benchmarks truly measure genuine grounded understanding of video content?* Recent evidence reveals vulnerabilities: high text-only performance [61] and substantial textual bias [31] suggest benchmarks may admit shortcuts, allowing models to answer correctly without deeply engaging visual evidence. This obscures genuine progress in video understanding or reasoning capabilities.

We identify that **temporal grounding in benchmarks**, requiring models to both answer correctly and localize relevant temporal segments, provides a principled diagnostic for video understanding. An answer obtained through world knowledge, language priors, or frame-level recognition need not identify when the supporting event occurs. Many multiple-choice benchmarks evaluate answer correctness alone, overlooking this distinction between answering and locating evidence. This gap motivates our investigation into benchmark quality and grounding-aware evaluation.

Unlike natural language understanding tasks where models process complete textual information, video question-answering faces an inherent challenge: **information loss from frame sampling**. Video content must be sampled into discrete, sparse frames due to context limits, and this sampling

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

discards temporal information. We adopt a unified methodological approach centered on controlling information loss through frame sampling density. By systematically varying input from black-video and endpoint controls through sparse (0.02 to 0.2 FPS) to denser (1 to 2 FPS) sampling, we measure how visual information availability impacts performance. This experimental apparatus, applied across matched reasoning modes on VideoMME [13], Video-MMMU [22], and LVBench [54], diagnoses shortcuts and tests how reasoning scales with information availability. Our complementary benchmark adds a temporal-grounding requirement to the same analysis.

Key findings. Our investigation reveals four complementary insights. First, **existing benchmarks contain widespread shortcuts**: model-response classification identifies 33.33% of VideoMME, 41.00% of Video-MMMU, and 12.46% of LVBench questions with at least one detected shortcut route (Figure 1). Second, **temporal grounding requirements expose an evaluation gap**: correct answers need not be accompanied by localized evidence, so answering and grounding must be evaluated separately. Third, **reasoning gains on existing benchmarks are larger on detected shortcut questions in aggregate**: the paired difference between shortcut and remaining-question gains is +6.83 points, with a 95% confidence interval of [+5.68, +7.98]. On TGVMME, reasoning effects under sparse sampling are small or mixed, while denser input permits larger gains whose magnitude depends on the model and grounding metric. Fourth, **input scaling provides larger gains in the Qwen comparison**: the 0.2-to-1 FPS step adds +19.84 answer points on average, compared with +1.44 points from enabling reasoning at 1 FPS. These interventions are not compute-matched, and the ordering is model-dependent rather than universal.

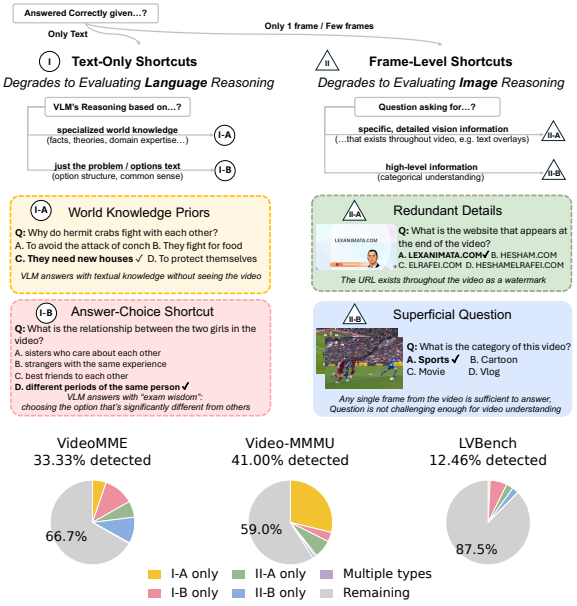


Figure 1: **When Video Benchmarks Test Everything But Videos.** Shortcut taxonomy across video benchmarks, illustrating knowledge, option structure, persistent details, and superficial recognition. The schematics depict solution routes; evaluated controls use black video and two endpoint frames. The pies count unique questions, separating multiple-category membership; Table 1 reports overlapping category totals.

Contributions. We deliver a four-type shortcut taxonomy with sampled human validation and a classifier analysis of three popular video benchmarks. TGVMME supplies a 1,500-question test bed that evaluates answers together with temporal evidence. Together, these components enable controlled comparisons of input scaling and inference-time reasoning across the Qwen3-VL and Gemma-4 families, connecting reasoning gains to both the supplied frames and the observed solution routes.

2 Related Work

2.1 Video Reasoning with RLVR

Following the success of reinforcement learning from verifiable rewards (RLVR) on mathematical and code reasoning [21, 23, 40], numerous works have applied RLVR to video understanding [28, 58], where video frames and questions are provided to VLMs for reasoning before answering. Recent efforts span supervised fine-tuning with chain-of-thought [7, 12, 41, 53, 83], reinforcement learning with temporal rewards [27, 47, 59], and test-time scaling through extended inference [36, 59, 62, 65, 77]. Video-R1 [12] introduces T-GRPO for temporal reasoning, Time-R1 [57] applies RL to temporal video grounding, and VideoChat-R1.5 [65] explores iterative perception. These approaches motivate studying how inference-time reasoning interacts with the visual evidence available to the model [46].

Our controlled comparison asks how information loss from sparse frame sampling constrains reasoning and how input scaling compares with enabling reasoning at fixed sampled frames. For Qwen3-VL, the 0.2-to-1 FPS input step adds +19.84 answer points, compared with +1.44 points from enabling reasoning at 1 FPS; this is an accuracy comparison, not a gain-per-compute estimate. CyberV [35], CoT-Vid/VISTA [24], and Rethinking Chain-of-Thought Reasoning for Videos [82] study strategies with model- and input-dependent reasoning effects. The separately named VISTA augmentation framework [38] addresses long-duration and high-resolution inputs; our analysis instead compares matched reasoning modes on fixed visual inputs and shortcut groups.

2.2 Evaluation of Video Understanding

Video understanding benchmarks have evolved toward sophisticated evaluation protocols. VideoMME [13] provides comprehensive assessment with perception, cognition, and temporal understanding questions. Video-MMMU [22] evaluates three cognitive stages using professional educational videos. LVBench [54] targets ultra-long videos with adversarial filtering against GPT-4 to reduce shortcuts. Earlier efforts including MVBench [26], EgoSchema [33], and TempCompass [30] established foundations for rigorous evaluation. Despite careful curation, recent evidence reveals persistent vulnerabilities: behavioral studies show unexpectedly high text-only performance [61], attention analyses reveal text-over-visual bias [31], and systematic audits identify benchmark vulnerabilities [74]. Creating high-quality video benchmarks requires substantial manual effort: unlike image labeling where single frames can be annotated quickly, video annotation requires watching entire sequences and crafting questions that probe genuine event-level comprehension [54].

2.3 Shortcuts in Vision and Language

Shortcut learning, exploiting superficial patterns rather than genuine understanding [10, 17], poses persistent challenges in multi-modal evaluation [63]. Vision-language models naturally inherit systematic textual bias, answering through linguistic conventions, answer-choice patterns, or memorized knowledge [71] from their underlying language model backbones [81]. For image understanding with VLMs, MMStar [6] identified evaluation samples that lack image dependency.

Prior video benchmarks filter language-answerable questions [54], while temporal-bias diagnostics and language-bias analyses further examine whether success depends on the intended evidence [9, 43, 68, 78]. Grounded VideoQA already evaluates answers together with temporal evidence [61]. Our contribution connects response-level shortcut classification to matched reasoning effects across visual budgets, using TGVMME as a complementary grounded test bed rather than claiming to introduce shortcut filtering or grounded VideoQA.

3 Identifying Shortcuts in Common Video Reasoning Benchmarks

An extreme diagnostic of the information-loss issue introduced in Section 1 is to ask whether video-language models can answer correctly when the source video is replaced by black frames. Success in this condition can reflect language priors or memorized world knowledge rather than the queried video event. We use this regime, together with sparse frame budgets, as a controlled probe for shortcut exploitation.

Methodology. We toggle configured thinking modes across the Qwen3-VL [3] and Gemma-4 [18] series while holding the checkpoint, question, options, and sampled frames fixed within each pair. Input settings progress from black-video and two-endpoint controls through sparse sampling (0.02 to 0.2 FPS) to denser sampling (1 to 2 FPS); Appendix D specifies the controls and preprocessing. Black video still supplies visual tokens, and Video-MMMU's Adaptation questions retain their associated static image. The sampler caps input at 768 frames and can reduce spatial resolution to fit its pixel budget; requested FPS is therefore not an equal-compute intervention. By examining successful Qwen3-VL responses under altered visual inputs and analyzing their reasoning traces, we identify apparent shortcut solution routes. We then classify potential shortcut problems with an LLM under a manually validated taxonomy, using the guidelines reported in Appendix C.

Table 1: **Shortcut annotations span text-based and frame-level routes.** Entries count unique question IDs classified from successful Qwen3-VL thinking responses; parentheses give percentages, and categories can overlap. The two TGVMME columns apply answer-only and answer-plus-positive-overlap success criteria before classification, measuring credited shortcut success rather than a change in intrinsic shortcut prevalence.

Category	VideoMME $n = 2{,}700$	Video-MMMU $n = 900$	LVBench $n = 1{,}549$	TGVMME $n = 1{,}500$	TGVMME + TG $n = 1{,}500$
I-A: World knowledge	145 (5.37%)	265 (29.44%)	12 (0.77%)	3 (0.20%)	0 (0.00%)
I-B: Answer-choice structure	317 (11.74%)	36 (4.00%)	101 (6.52%)	67 (4.47%)	18 (1.20%)
II-A: Persistent details	171 (6.33%)	70 (7.78%)	42 (2.71%)	69 (4.60%)	26 (1.73%)
II-B: Superficial recognition	275 (10.19%)	7 (0.78%)	39 (2.52%)	16 (1.07%)	8 (0.53%)
Any category (union)	900 (33.33%)	369 (41.00%)	193 (12.46%)	149 (9.93%)	49 (3.27%)

3.1 A Taxonomy of Shortcuts

Since video reasoning models produce chain-of-thought text before their answer, their traces can suggest whether the answer relies on visual evidence, textual priors, or option structure, although individual traces may contain hallucinations. These categories describe observable response patterns rather than verified internal mechanisms.

Our manual examination reveals two broad categories of shortcuts that allow models to bypass genuine video understanding (Figure 1). **Text-Only Shortcuts** enable answering questions with zero visual input: (I-A) *World Knowledge Priors* exploit specialized domain knowledge, facts, or theories memorized during pretraining, and (I-B) *Answer-Choice Shortcuts* leverage linguistic patterns in question or option structure. **Frame-Level Shortcuts** degrade video understanding to image understanding, solvable with minimal visual evidence: (II-A) *Redundant Details* ask about persistent visual elements appearing throughout the video (e.g., watermarks, text overlays, or static scene properties visible in any frame), and (II-B) *Superficial Questions* require only high-level or categorical understanding of video content, answerable from any single frame (e.g., categorizing a video as "sports" from one frame showing a soccer field).

To validate this taxonomy at scale, we developed objective decision guidelines and employed GPT-5 with reasoning capability to classify questions; the decision tree is visualized in Figure 1. For each question where models answered correctly with no or minimal visual input, we provided the question text, answer options, correct answer, and model reasoning trace. The classifier decides whether the response relies on world knowledge (I-A), answer-choice bias (I-B), persistent visual details (II-A), shallow categorical recognition (II-B), or no systematic shortcut. Section C.1 reports human validation of this LLM-as-a-judge setup.

3.2 Widespread Shortcuts Across Benchmarks

Table 1 reports shortcut distributions within three widely used video understanding benchmarks. A substantial portion of questions admit language-prior or frame-level solution routes without requiring temporal analysis. Deduplicating question identifiers within categories and taking their union gives 33.33% on VideoMME, 41.00% on Video-MMMU, and 12.46% on LVBench; categories can overlap. World knowledge accounts for 29.44% of Video-MMMU questions. Even LVBench, which filters questions answered correctly by both GPT-4 and GLM-4 without video, retains detected text-based and frame-level shortcuts.

For the reasoning comparison, we hold this stratification fixed and score each group directly. The evaluated shortcut groups contain 899 VideoMME, 369 Video-MMMU, and 193 LVBench questions; one VideoMME annotation identifier does not map to an evaluated key. The complement, denoted Non-SC, means no detected shortcut rather than intrinsic shortcut freedom. Selecting labels from successful thinking responses favors thinking under the labeling conditions, so the group contrast is descriptive, not a causal effect of shortcut membership.

These findings expose fundamental limitations of the multiple-choice format for evaluating video understanding. When models can answer correctly through language priors, world knowledge, or single-frame analysis, impressive accuracy numbers may reflect everything *except* genuine temporal

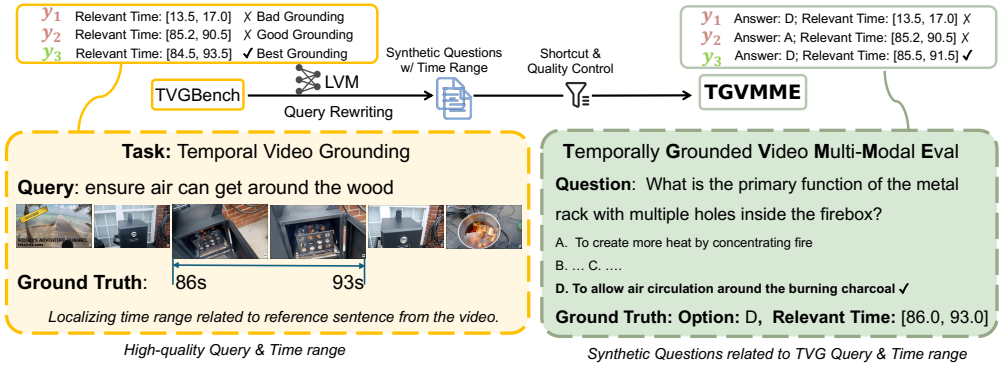


Figure 2: **TGVMME construction pipeline.** Temporal video grounding annotations from TVG-Bench are expanded into multiple-choice questions paired with their reference time ranges through video-conditioned question generation and quality control. The query-rewriting stage denotes semantic expansion conditioned on the video, source query, and reference interval, rather than lexical paraphrasing alone.

video reasoning. This motivates our development of more grounded evaluation protocols that require models to demonstrate not just correct answers, but verifiable evidence of having comprehended the video content through temporal localization, ensuring evaluation measures what we intend to measure rather than rewarding sophisticated shortcut exploitation.

4 TGVMME: Benchmarking Video Understanding with Temporal Grounding

Having identified widespread shortcuts in existing benchmarks, we now introduce TGVMME (**T**emporally **G**rounded **V**ideo **M**ulti-**M**odal **E**valuation), a benchmark designed to assess video understanding through temporal grounding. Unlike benchmarks that only evaluate multiple-choice answer correctness, TGVMME requires models to select the correct answer *and* localize the relevant temporal segment that supports it. We construct TGVMME from human temporal video grounding annotations and transform them into multiple-choice questions with associated time ranges (Figure 2); evaluation reports both standard answer accuracy and grounded accuracy.

4.1 Benchmark Construction and Curation

TGVMME is constructed with temporal video grounding (TVG) annotations from TVGBench [57], which provide natural language queries paired with human-annotated temporal boundaries indicating when the queried event occurs in the video. We transform these TVG annotations into multiple-choice questions through a controlled, video-conditioned generation process (Figure 2).

Question expansion and option generation. Starting from a TVG query (e.g., "ensure air can get around the wood") and its reference time range, Qwen3-VL-32B-Thinking receives the video, query, and interval and generates a question, correct answer, and distractors. The question can introduce details supported by the video rather than merely rephrase the source sentence. The resulting item inherits the source temporal interval, which provides the reference for grounding evaluation.

Quality control and shortcut filtering. We apply structural checks and behavioral filters based on Qwen3-VL-2B and 8B answer correctness under altered visual inputs. The final shortcut-filtering stage removes 240 candidates from a 1,740-item pool, leaving 1,500 multiple-choice questions. These checks screen answerability patterns; they do not independently establish that every question is unanswerable from frames outside its annotated interval. Appendix E specifies the construction roles and filtering procedure.

A human annotator also checked 150 sampled items for video relevance, answer-choice consistency, and answer correctness; 139 passed all checks. The reference intervals originate from human TVG annotations, but this item-level validation did not separately annotate entailment of every generated

detail within the inherited segment. Appendix C.3 describes this scope and the sampled failure modes.

4.2 Evaluation Protocol: Grounded Accuracy

We introduce **Temporally Grounded Accuracy** as our primary metric, requiring the correct answer option and positive overlap between the predicted and reference temporal intervals. This dual requirement distinguishes answer correctness from success in locating supporting evidence. Positive overlap is permissive: a broad interval can satisfy it without precise localization.

We therefore also report answer-correct tIoU@0.5 and IoP@0.5, with absolute mean overlaps and paired confidence intervals in Appendix F. tIoU divides intersection duration by union duration, whereas IoP divides it by predicted duration; all-question mean overlaps include incorrect answers, with invalid intervals assigned zero. Standard multiple-choice accuracy serves as the reference point, measuring answer correctness without temporal grounding. The gap between answer accuracy and grounded accuracy is a localization gap, not by itself an estimate of shortcut prevalence. All matched TGVMME results use the same 1,500 questions and fixed denominators, with malformed answers scored incorrect.

4.3 Cross-Benchmark Validation of TGVMME Quality

To assess TGVMME alongside existing benchmarks, we apply the shortcut taxonomy used in Table 1. The three existing benchmarks contain 33.33%, 41.00%, and 12.46% unique shortcut-labeled questions, respectively, while TGVMME retains 149 of 1,500 questions with a detected shortcut label (9.93%). Because the construction filter uses related shortcut labels, this difference describes the curated test bed rather than an independent estimate of filtering generalization.

Requiring positive temporal overlap before shortcut classification yields 49 labeled questions (3.27%), compared with 149 (9.93%) under answer-only success. These separately classified sets measure shortcut-enabled credited success under the respective criteria, not a change in intrinsic shortcut prevalence or a paired removal count. The answer and grounded scores in Table 2, supplemented by stricter localization metrics, measure this distinction directly. We next use this evaluation to analyze how reasoning benefits change with the available frames.

5 Analyzing Reasoning on Videos

Using TGVMME, we now conduct systematic experiments to investigate whether inference-time reasoning improves video understanding, under what conditions, and how its benefit compares with increasing input tokens (frames). Our findings indicate that, under the evaluated sparse-sampling regimes, performance is constrained by available visual evidence, with the 0.2-to-1 FPS step providing more answer-accuracy gain than enabling reasoning for Qwen3-VL. All matched TGVMME results use the same 1,500 questions; the comparisons are not compute-matched.

5.1 Inference-Time Reasoning Benefits Scale With Frame Budget

Reusing the controlled probe of Section 3 (Qwen3-VL and Gemma-4 with thinking toggled on and off across black-video, minimum-input, sparse 0.02-to-0.2 FPS, and denser 1-to-2 FPS regimes), we now turn from *which questions are shortcut-solvable* to *whether reasoning helps when temporal evidence is required*. Sparse FPS is relevant to fixed-frame evaluation: a fixed frame budget on a long video can supply little temporal coverage even when every supplied frame is processed. Figure 3 and Appendix A describe the visual-budget constraints, distinguishing uncapped token projections from the realized capped inputs.

Reasoning helps when sufficient evidence is given and yields small or mixed effects otherwise. Table 2 reveals sparse and dense regimes for the five displayed models. Under sparse sampling the effect of reasoning is small in absolute terms: *reasoning yields small, mixed-direction effects when frame sampling is sparse*, with Gemma-4 retaining a positive answer lift while Qwen3-VL drifts toward neutral or negative, especially for the largest Qwen3-VL model. At 0.02 FPS, 2B drops by -3.20 answer points, 4B retains a +0.67-point lift, and 8B falls by -4.47 points. Grounded accuracy is more demanding for the 8B variant, with reasoning costing -3.00 TG points at 0.02 FPS.

On language tasks, longer chains of thought can help even with short input by drawing on latent parametric knowledge [21, 23, 44]. Video reasoning additionally depends on sampled evidence: a long trace cannot directly observe the temporal order of events from frames the model never received. At sparse FPS, generic priors and answer-elimination heuristics can support an answer without providing its temporal evidence. When those priors align with the correct answer, accuracy can be preserved without accurate localization.

Both families benefit from reasoning when frames are sufficient. The picture changes at higher frame rates. For Qwen3-VL at 1 FPS, reasoning adds small MCQ gains across sizes, and at 2 FPS the gains grow to +2.60 (2B), +4.40 (4B), and +0.47 (8B) points, with larger positive-overlap grounded gains. Gemma-4-E2B and E4B follow the same qualitative trajectory in answer accuracy. At sparse FPS, reasoning already improves Gemma-4 MCQ accuracy across both sizes, in contrast to the mixed Qwen3-VL effects. The gap widens with more frames: at 1 FPS reasoning yields +8.13 points (E2B) and +6.27 points (E4B) on MCQ, with grounded gains of +7.47 and +15.33 points. The sparse-to-dense contrast holds for these models, suggesting that reasoning utility depends jointly on the available frames and on the underlying model's ability to integrate evidence across them. Gemma-4-E2B and E4B peak in answer accuracy at 1 FPS rather than 2 FPS; this observation does not by itself establish a causal explanation from pretraining frame rates.

Table 2: **Reasoning benefits depend on frame budget.** Thinking-mode multiple-choice (MCQ) and positive-overlap grounded accuracy (TG, %) on 1,500 TGVMME questions; Δ is thinking minus no-thinking (absolute points). Bold marks within-model column maxima over the displayed FPS settings.

FPS	MCQ	Δ	TG	Δ
Qwen3-VL-2B				
0.02	21.40	−3.20	6.87	−1.20
0.2	38.53	+0.87	18.47	+1.87
1	55.80	+1.87	44.00	+4.33
2	**58.47**	**+2.60**	**48.07**	**+5.47**
Qwen3-VL-4B				
0.02	28.33	+0.67	8.73	−2.13
0.2	44.27	+0.40	23.40	+0.27
1	63.93	+2.33	51.07	+1.67
2	**68.20**	**+4.40**	**57.07**	**+4.00**
Qwen3-VL-8B				
0.02	24.13	−4.47	7.33	−3.00
0.2	42.20	−1.67	25.47	−0.27
1	64.80	+0.13	54.67	+1.87
2	**67.27**	**+0.47**	**57.47**	**+2.20**
Gemma-4-E2B				
0.02	26.87	+1.80	14.00	+2.60
0.2	40.13	+4.60	15.07	+4.93
1	**45.67**	**+8.13**	**20.53**	+7.47
2	44.20	+7.27	19.67	**+9.13**
Gemma-4-E4B				
0.02	32.00	+3.60	13.13	+2.73
0.2	45.80	+2.20	22.27	+7.20
1	**56.00**	**+6.27**	**32.53**	**+15.33**
2	53.47	+3.60	30.13	+13.93

Stricter metrics distinguish positive overlap from precise localization. At 0.02 FPS, none of the six evaluated Qwen3-VL and Gemma-4 models has a positive answer-correct tIoU@0.5 or IoP@0.5 gain whose paired 95% interval excludes zero; at 2 FPS, four improve both metrics with intervals excluding zero. Appendix F includes the full scores, 10,000-resample paired question-bootstrap intervals, and Gemma-4-31B boundary case, whose 1 FPS joint IoP@0.5 change is −2.9 points. These exploratory intervals are not adjusted for multiple comparisons and do not measure repeated-run variability.

Visual evidence as a constraint. These results indicate that, in the evaluated sparse-sampling regime, information available in the sampled frames constrains accuracy and reasoning provides only limited compensation. Unlike language reasoning tasks where extended thinking can surface latent knowledge from the model's parameters [21, 23], *reasoning cannot directly observe information that frame sampling has discarded*. This distinction motivates evaluating temporal localization separately from answer correctness and comparing additional frames with longer reasoning chains.

5.2 Input Scaling Outperforms Output Scaling for Qwen3-VL

Having established that reasoning depends on frame availability, we now directly compare increasing input tokens (frames) with enabling output-side reasoning. We evaluate performance at different frame rates with and without reasoning enabled; these accuracy comparisons do not estimate marginal benefit per token or equal-compute efficiency.

Table 3: **Input and thinking gains depend on density.** Absolute MCQ gains (points) on 1,500 TGVMME questions include the unweighted Qwen3-VL mean; bold marks the larger gain in each pair. Frame steps keep thinking enabled, whereas mode switches hold FPS fixed; these are not matched-compute comparisons.

Change	2B	4B	8B	Mean
Black video → 0.2 FPS				
Frames (thinking on)	**+15.27**	**+19.20**	**+16.93**	**+17.13**
Thinking at 0.2	+0.87	+0.40	−1.67	−0.13
0.2 FPS → 1 FPS				
Frames (thinking on)	**+17.27**	**+19.67**	**+22.60**	**+19.84**
Thinking at 1	+1.87	+2.33	+0.13	+1.44
1 FPS → 2 FPS				
Frames (thinking on)	**+2.67**	+4.27	**+2.47**	**+3.13**
Thinking at 2	+2.60	**+4.40**	+0.47	+2.49

Input scaling provides substantially larger improvements in the sparse-to-dense Qwen comparison. Table 3 quantifies the performance gain from input scaling (increasing FPS) versus output scaling (enabling reasoning) on the same Qwen3-VL-Thinking runs as Table 2. Moving from black video to sparse frame-rate (0.2 FPS) yields +17.13 points of family-averaged MCQ improvement, and moving from 0.2 to 1 FPS yields +19.84 points. In contrast, output scaling through reasoning changes accuracy by -0.13 points at 0.2 FPS, +1.44 points at 1 FPS, and +2.49 points at 2 FPS.

For these Qwen3-VL settings, *the sparse-to-dense input step provides a larger gain than enabling reasoning*. From 0.2 to 1 FPS, adding frames yields +19.84 average MCQ points versus +1.44 points from enabling reasoning at 1 FPS. Even at 2 FPS, where the family-average reasoning gain is largest, the preceding 0.2-to-1 FPS input step provides more benefit than reasoning's contribution.

Scaling from 1 to 2 FPS yields a smaller +3.13-point family-average improvement, showing that the marginal benefit of frames depends on the starting density. The ordering is also model-dependent: Gemma-4-E2B gains +5.5 MCQ points from 0.2 to 1 FPS, compared with +8.1 points from enabling reasoning at 1 FPS. For compute cost, we report output lengths and average running times per question across model sizes, FPS, and reasoning settings in Appendix C.4; these descriptive profiles do not establish compute-normalized superiority.

5.3 Reasoning Gains on Existing Benchmarks Are Concentrated on Shortcuts

Having shown that reasoning's benefit on TGVMME depends on frame budget, we now examine how reasoning gains are distributed on existing benchmarks. We evaluate the same models with and without reasoning on shortcut versus remaining questions across VideoMME, Video-MMMU, and LVBench, asking whether gains are concentrated on shortcut-solvable items.

Table 4 reveals a predominant aggregate pattern: reasoning improves performance more on detected shortcut questions, although individual conditions have mixed directions. Thinking increases shortcut accuracy in 28 of 36 benchmark-by-size-by-input settings and remaining-question accuracy in 15. On VideoMME, the 2B and 4B models improve shortcut accuracy by +1.8 to +11.9 points across different visual input conditions, while remaining-question deltas range from -4.8 to +0.9 points. Video-MMMU exhibits a larger asymmetry: for the 2B model, reasoning improves shortcuts by +11.1 to +26.6 points, while remaining-question changes range from -13.7 to +2.6 points across sampling densities. The 8B model is an important exception on VideoMME and LVBench, losing accuracy on both groups in all four displayed conditions.

Explaining apparent reasoning gains on existing benchmarks. This asymmetry helps interpret reasoning gains on VideoMME and Video-MMMU. Table 1 identifies 33.33% and 41.00% shortcut-labeled questions in these benchmarks, respectively. We average each question's paired effect over the three model sizes and four input conditions, then compare question-weighted group means. The 1,461 shortcut questions gain +3.99 points, versus -2.84 points for the 3,688 remaining questions: a +6.83-point contrast with a paired 95% interval of [+5.68, +7.98]. This comparison includes

Table 4: **Thinking effects differ across question groups.** Qwen3-VL accuracy (%) is shown with thinking disabled (NR) and enabled (R), with absolute-point deltas and bold marking the better NR/R value in each pair. Endpoints uses two endpoint frames; Video-MMMU Black video retains the static image for Adaptation questions.

Benchmark	Size	Mode	Black video		Endpoints		0.02 FPS		0.2 FPS	
			SC	Non-SC	SC	Non-SC	SC	Non-SC	SC	Non-SC
VideoMME	2B	NR	41.60	20.77	59.73	**20.93**	63.85	**26.04**	78.31	38.09
		R	**43.38**	**21.65**	**68.97**	16.27	**66.63**	24.82	**85.76**	**38.59**
		Δ	+1.78	+0.89	+9.23	−4.66	+2.78	−1.22	+7.45	+0.50
	4B	NR	45.94	**18.88**	67.63	**20.71**	74.30	**28.76**	86.87	45.75
		R	**48.83**	18.10	**79.53**	15.94	**77.75**	28.60	**93.55**	**45.81**
		Δ	+2.89	−0.78	+11.90	−4.78	+3.45	−0.17	+6.67	+0.06
	8B	NR	**50.95**	**19.77**	**69.52**	**21.15**	**73.64**	**27.93**	**86.43**	**44.42**
		R	40.71	14.88	58.84	10.77	64.29	20.49	80.09	39.70
		Δ	−10.23	−4.89	−10.68	−10.38	−9.34	−7.44	−6.34	−4.72
Video-MMMU	2B	NR	41.19	**21.28**	45.26	**21.09**	45.80	**28.63**	48.24	30.32
		R	**52.30**	9.60	**57.18**	7.34	**68.83**	24.29	**74.80**	**32.96**
		Δ	+11.11	−11.68	+11.92	−13.75	+23.04	−4.33	+26.56	+2.64
	4B	NR	52.30	12.24	60.70	**9.60**	67.75	**27.50**	69.65	34.09
		R	**62.87**	**12.81**	**65.85**	6.21	**71.82**	24.48	**83.47**	**36.91**
		Δ	+10.57	+0.56	+5.15	−3.39	+4.07	−3.01	+13.82	+2.82
	8B	NR	58.54	**22.22**	61.52	**24.67**	69.38	32.20	76.15	40.30
		R	**71.82**	20.15	**81.84**	11.68	**84.82**	**36.35**	**91.60**	**44.07**
		Δ	+13.28	−2.07	+20.33	−12.99	+15.45	+4.14	+15.45	+3.77
LVBench	2B	NR	50.26	18.58	59.59	**21.98**	68.91	29.94	76.17	39.23
		R	**54.40**	**19.25**	**76.17**	20.21	**81.35**	**31.71**	**88.08**	**41.15**
		Δ	+4.15	+0.66	+16.58	−1.77	+12.44	+1.77	+11.92	+1.92
	4B	NR	59.07	20.35	66.32	19.69	77.72	30.97	81.87	40.63
		R	**60.10**	**21.90**	**80.83**	**21.09**	**88.08**	**34.51**	**88.60**	**43.58**
		Δ	+1.04	+1.55	+14.51	+1.40	+10.36	+3.54	+6.74	+2.95
	8B	NR	**61.66**	**22.05**	**78.24**	**19.54**	**72.02**	**31.19**	**83.94**	**45.28**
		R	40.93	14.31	50.26	11.65	49.22	16.45	69.43	35.62
		Δ	−20.73	−7.74	−27.98	−7.89	−22.80	−14.75	−14.51	−9.66

mode-dependent answer validity, with invalid outputs retained as incorrect rather than removed from the denominator. Thus benchmark composition matters when interpreting aggregate reasoning gains, without implying that every benchmark or cell follows the same pattern.

At denser 2 FPS input on VideoMME and Video-MMMU, shortcut gains exceed remaining-question gains in five of six matched settings (Appendix F.3). The same pooled labels are used, and their exact source FPS is not recorded, so this extension is not guaranteed to be outside the labeling conditions.

Validating TGVMME through grounded comparison. Our benchmark provides a complementary condition in which correct answers must also receive temporal-grounding credit. Section 5.1 showed that answer gains on TGVMME grow with the frame budget for the five main-table models, with family-averaged gains of +2.49 points for Qwen3-VL at 2 FPS and +7.20 points for Gemma-4-E2B/E4B at 1 FPS. These gains must nevertheless be distinguished from localization changes: Gemma-4-E2B at 1 FPS gains +8.13 answer points but +0.80 tIoU@0.5 and +3.27 IoP@0.5 points. Grounded evaluation therefore tests a requirement that answer accuracy alone does not establish.

Behavioral interpretation. Under sparse sampling, textual cues can supply a correct answer while visual evidence remains fragmentary. The observed group contrast concerns success-selected labels and configured thinking modes, not a causal effect of shortcut membership or a verified internal mechanism. *In the evaluated matched settings, reasoning gains favor detected shortcut questions in aggregate, while reliable temporal evidence remains a separate requirement.*

6 Conclusion

We investigated whether existing video benchmarks truly measure genuine grounded understanding. By systematically controlling frame sampling and analyzing models with and without reasoning, we revealed that existing benchmarks contain substantial shortcuts that let models bypass temporal video understanding, and that requiring temporal grounding distinguishes correct answers from located evidence. To address these limitations, we introduced TGVMME, a temporally grounded benchmark with 1,500 questions requiring both correct answers and temporal localization, together with a shortcut taxonomy and sampled human validation.

Using TGVMME to analyze reasoning models, we found that inference-time reasoning yields gains that grow with the frame budget across the five main-table models (+2.49 answer points across Qwen3-VL at 2 FPS and +7.20 points for Gemma-4-E2B/E4B at 1 FPS, with small or mixed effects at sparse sampling). The reasoning gains measured on existing benchmarks are larger on detected shortcut questions in aggregate, with a +6.83-point paired contrast. Input scaling delivers larger gains for Qwen3-VL, with the 0.2-to-1 FPS step yielding +19.84 average MCQ points compared with +1.44 points from enabling reasoning; this ordering is model-dependent.

Our results highlight the importance of grounded evaluation: *unlike language reasoning, where extended thinking can surface latent parametric knowledge, reasoning cannot directly observe information that frame sampling has discarded*. Future work should prioritize frame-efficient encoding and temporal localization alongside reasoning, and gains attributed to test-time reasoning should be reported alongside the visual-input budget. Our matched comparisons concern configured thinking modes at fixed sampled frames; Appendix C.6 discusses the scope of the shortcut labels and human validation.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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

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

[5] Kehua Chen. Chronoforge-rl: Chronological forging through reinforcement learning for enhanced video understanding. *arXiv preprint arXiv:2509.15800*, 2025.

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

[7] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[9] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for VideoLLMs. In *British Machine Vision Conference (BMVC)*, 2025.

[10] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1): 110–120, 2023.

[11] Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025.

[12] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[14] Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

[15] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[16] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025.

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[18] Google DeepMind. Gemma 4 Model Card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed: 2026-05-06.

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[20] Chaohong Guo, Xun Mo, Yongwei Nie, Xuemiao Xu, Chao Xu, Fei Yu, and Chengjiang Long. Tar-tvg: Enhancing vlms with timestamp anchor-constrained reasoning for temporal video grounding. *arXiv preprint arXiv:2508.07683*, 2025.

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[22] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[23] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[24] Hongbo Jin, Jiayu Ding, Siyi Xie, Guibo Luo, and Ge Li. VISTA: Mitigating semantic inertia in Video-LLMs via training-free dynamic chain-of-thought routing. *arXiv preprint arXiv:2505.11830*, 2026. Version 3, revised January 2026; first posted May 2025 as CoT-Vid.

[25] Daeun Lee, Jaehong Yoon, Jaemin Cho, and Mohit Bansal. Video-skill-cot: Skill-based chain-of-thoughts for domain-adaptive video reasoning. *arXiv preprint arXiv:2506.03525*, 2025.

[26] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[27] Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. Tempsamp-r1: Effective temporal sampling with reinforcement fine-tuning for video llms. *arXiv preprint arXiv:2509.18056*, 2025.

[28] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, et al. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *arXiv preprint arXiv:2505.04921*, 2025.

[29] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[30] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

[31] Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, et al. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9087–9097, 2025.

[32] Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407, 2025.

[33] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[34] Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3 video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

[35] Jiahao Meng, Shuyang Sun, Yue Tan, Lu Qi, Yunhai Tong, Xiangtai Li, and Longyin Wen. CyberV: Cybernetics for test-time scaling in video understanding. *arXiv preprint arXiv:2506.07971*, 2025.

[36] Jinyoung Park, Jeehye Na, Jinyoung Kim, and Hyunwoo J Kim. Deepvideo-r1: Video reinforcement fine-tuning via difficulty-aware regressive grpo. *arXiv preprint arXiv:2506.07464*, 2025.

[37] Yi Peng, Peiyu Wang, Xiaokun Wang, Yichen Wei, Jiangbo Pei, Weijie Qiu, Ai Jian, Yunzhuo Hao, Jiachun Pan, Tianyidan Xie, et al. Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. *arXiv preprint arXiv:2504.05599*, 2025.

[38] Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. VISTA: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[39] Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv preprint arXiv:2504.12083*, 2025.

[40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[41] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

[42] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024.

[43] Nina Shvetsova, Arsha Nagrani, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Unbiasing through textual descriptions: Mitigating representation bias in video benchmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[44] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[45] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

[46] Yunlong Tang, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, et al. Video-lmm post-training: A deep dive into video reasoning with large multimodal models. *arXiv preprint arXiv:2510.05034*, 2025.

[47] Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, et al. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*, 2025.

[48] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[49] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[50] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[51] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

[52] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[53] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025.

[54] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024.

[55] Xiao Wang, Liye Jin, Xufeng Lou, Shiao Wang, Lan Chen, Bo Jiang, and Zhipeng Zhang. Reasoningtrack: Chain-of-thought reasoning for long-term vision-language tracking. *arXiv preprint arXiv:2508.05221*, 2025.

[56] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*, 2023.

[57] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-r1: Post-training large vision language model for temporal video grounding. *arXiv preprint arXiv:2503.13377*, 2025.

[58] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

[59] Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-rts: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28114–28128, 2025.

[60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[61] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.

[62] Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025.

[63] Zhuo Xu, Xiang Xiang, and Yifan Liang. Overcoming shortcut problem in vlm for robust out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15402–15412, 2025.

[64] Shuhang Xun, Sicheng Tao, Jungang Li, Yibo Shi, Zhixin Lin, Zhanhui Zhu, Yibo Yan, Hanqian Li, Linghao Zhang, Shikang Wang, et al. Rtv-bench: Benchmarking mllm continuous perception, understanding and reasoning through real-time video. *arXiv preprint arXiv:2505.02064*, 2025.

[65] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

[66] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[67] Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.

[68] Yiming Yang, Yangyang Guo, Hui Lu, and Yan Wang. VidLBEval: Benchmarking and mitigating language bias in video-involved LVLMs. *arXiv preprint arXiv:2502.16602*, 2025.

[69] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[70] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[71] Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Wei Ye, Jindong Wang, Xing Xie, Yue Zhang, and Shikun Zhang. Kieval: A knowledge-grounded interactive evaluation framework for large language models. *arXiv preprint arXiv:2402.15043*, 2024.

[72] Zhuohao Yu, Chang Gao, Wenjin Yao, Yidong Wang, Zhengran Zeng, Wei Ye, Jindong Wang, Yue Zhang, and Shikun Zhang. Freeeval: A modular framework for trustworthy and efficient evaluation of large language models. *arXiv preprint arXiv:2404.06003*, 2024.

[73] Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. Rewardanything: Generalizable principle-following reward models. *arXiv preprint arXiv:2506.03637*, 2025.

[74] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

[75] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

[76] Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *arXiv preprint arXiv:2503.18013*, 2025.

[77] Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. *arXiv preprint arXiv:2508.04416*, 2025.

[78] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

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

[80] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636, 2025.

[81] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*, 2023.

[82] Yiwu Zhong, Zi-Yuan Hu, Yin Li, and Liwei Wang. Rethinking chain-of-thought reasoning for videos. *arXiv preprint arXiv:2512.09616*, 2025.

[83] Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. Reagent-v: A reward-driven multi-agent framework for video understanding. *arXiv preprint arXiv:2506.01300*, 2025.

[84] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025.

A Visual Budgets for Long-Form Video

Frame density and thinking change different parts of the inference workload. The profiles in this section provide descriptive context for Table 3, not equal-compute performance comparisons.

A.1 Frame Counts and Token Budgets

For a fixed per-frame token allocation, uncapped visual-token demand grows with video duration and sampling density. Figure 3 illustrates this relationship by extrapolating from a 0.1 FPS reference. Its projected 1 FPS costs are approximately 182K, 340K, and 932K tokens for Video-MMMU, VideoMME, and LVBench, respectively. These projections describe an uncapped constant-resolution scenario, not the realized token lengths in our capped, resolution-adaptive evaluation.

Long videos can therefore reach a frame or token budget before achieving the requested temporal density. The practical trade-off depends on duration, spatial resolution, architecture, batching, and hardware. Standard dense attention can incur quadratic prefill work in sequence length, whereas KV-cache storage grows linearly at fixed model dimensions. Neither relationship alone determines measured end-to-end latency or makes a particular FPS universally representative of deployment.

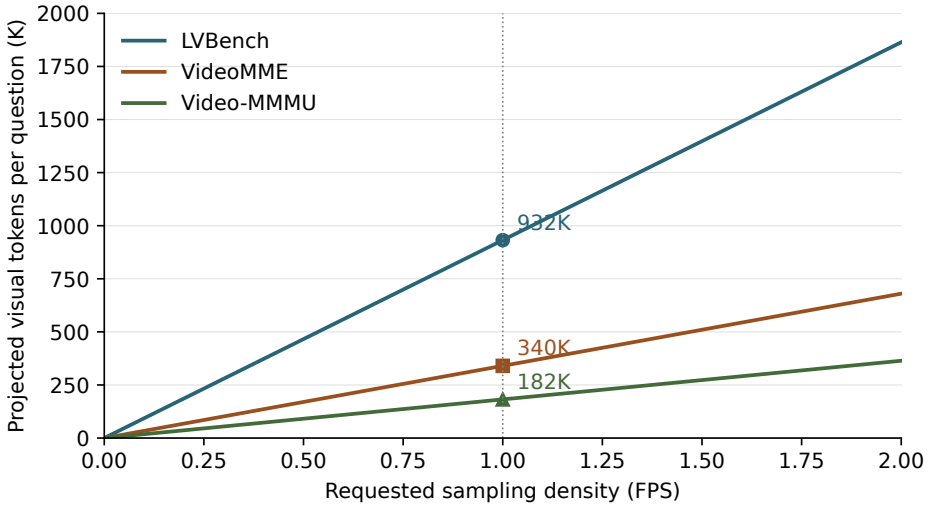


Figure 3: **Uncapped token demand illustrates the pressure of long videos on visual budgets.** The curves extrapolate fixed per-frame costs from a 0.1 FPS reference and mark projected costs at 1 FPS. They illustrate duration-dependent input demand rather than the realized token counts of the capped, resolution-adaptive experiments.

B Additional TGVMME Results

Due to space limitations, the main text reports the Qwen3-VL and Gemma-4 series most prominently. Table 5 provides the Qwen3-VL results alongside Qwen2.5-VL, Video-R1, and Open-O3-Video, distinguishing answer accuracy from positive-overlap grounded accuracy. Video-R1 and Open-O3-Video have no matched no-thinking controls here, so their absolute scores characterize performance at different input budgets rather than the causal effect of enabling reasoning.

Open-O3-Video improves grounded accuracy from 11.47% under black-video input to 18.60% at 2 FPS, a +7.13-point difference over the same 1,500 questions. Video-R1 improves from 9.87% to 11.93% between black-video input and 0.2 FPS, a +2.06-point difference; its 2 FPS scores describe only the 128 available questions and are not a matched full-benchmark comparison. The paired Qwen3-VL and Gemma-4 comparisons, strict localization metrics, and uncertainty estimates appear in Appendix F.

Table 5: **Answer and grounded accuracy across additional video models.** R and NR denote reasoning and no-reasoning conditions; TG requires a correct answer and positive temporal overlap. All rows use 1,500 questions except Video-R1 at 2 FPS (†), which reports the 128 available questions and is not directly comparable to full-set rows; bold marks the best full-set score within each model and score column.

Model	Input	N	MCQ R	MCQ NR	TG R	TG NR
Qwen3-VL-2B	Black	1500	23.27	21.73	7.07	6.93
	Min	1500	21.73	24.00	7.00	7.73
	0.02	1500	21.40	24.60	6.87	8.07
	0.2	1500	38.53	37.67	18.47	16.60
	1	1500	55.80	53.93	44.00	39.67
	2	1500	**58.47**	**55.87**	**48.07**	**42.60**
Qwen3-VL-4B	Black	1500	25.07	24.60	8.27	8.20
	Min	1500	29.00	28.00	9.20	11.33
	0.02	1500	28.33	27.67	8.73	10.87
	0.2	1500	44.27	43.87	23.40	23.13
	1	1500	63.93	61.60	51.07	49.40
	2	1500	**68.20**	**63.80**	**57.07**	**53.07**
Qwen3-VL-8B	Black	1500	25.27	26.67	7.93	8.53
	Min	1500	23.73	27.73	6.93	9.27
	0.02	1500	24.13	28.60	7.33	10.33
	0.2	1500	42.20	43.87	25.47	25.73
	1	1500	64.80	64.67	54.67	52.80
	2	1500	**67.27**	**66.80**	**57.47**	**55.27**
Video-R1-7B	Black	1500	26.13	—	9.87	—
	Min	1500	27.27	—	9.47	—
	0.02	1500	28.80	—	10.20	—
	0.2	1500	**32.53**	—	**11.93**	—
	2	128†	40.62	—	21.88	—
Qwen2.5-VL-7B	Black	1500	10.60	23.27	2.67	5.40
	Min	1500	24.87	26.87	8.73	8.93
	0.02	1500	25.67	26.40	9.33	9.00
	0.2	1500	34.67	33.20	15.53	15.40
	2	1500	**40.73**	**39.80**	**21.93**	**17.80**
Open-O3-Video	Black	1500	30.00	—	11.47	—
	Min	1500	29.80	—	11.73	—
	0.02	1500	30.60	—	13.20	—
	0.2	1500	**36.13**	—	16.33	—
	2	1500	35.20	—	**18.60**	—

Table 6: **Human inspection supports the sampled positive shortcut assignments.** A single annotator checked 50 assignments per category; the rates measure acceptance of these labels, not recall or inter-annotator agreement.

Shortcut Type	Valid / Total	Invalid	Validity Rate
I-A: World knowledge	48 / 50	2	96.00%
I-B: Answer-choice structure	47 / 50	3	94.00%
II-A: Persistent details	50 / 50	0	100.00%
II-B: Superficial recognition	47 / 50	3	94.00%
Overall	192 / 200	8	96.00%

C Shortcut Classification Guidelines

The classifier receives a question identifier, question, choices, labeled answer, and successful model responses under removed or sparse video input. It does not receive the video itself. Its task is to identify a coherent apparent route to the answer, returning I-A, I-B, II-A, II-B, or N for insufficient shortcut evidence, together with a brief explanation.

World-knowledge shortcuts (I-A) use substantive factual or conceptual knowledge that can supply the answer without the event being observed. Answer-choice shortcuts (I-B) use option structure, common-sense elimination, or test-taking cues. Persistent-detail shortcuts (II-A) rely on specific visual attributes such as an overlay, color, or static numerical detail. Superficial-recognition shortcuts (II-B) rely on a broad category rather than the video's temporal development. These categories describe the observed responses, not mutually exclusive intrinsic properties of questions.

The decision process first inspects a correct removed-video response for a coherent knowledge-based or option-based justification. A lucky guess, unsupported observation, or visually hallucinated explanation is not sufficient evidence of a systematic text-based route. Vision-enabled responses are then considered for persistent details or broad recognition. Category assignments can differ across responses to the same question, so Table 1 deduplicates within each category and separately reports their union. The labeling inputs are grouped as removed-video, minimum-frame, and higher-density responses rather than by exact FPS. The 2 FPS analysis uses this same pooled stratification at an additional input condition. Selection on successful thinking responses favors thinking within the labeling conditions, so the group contrast is descriptive rather than a causal effect of shortcut membership.

For TGVMME, the answer-only probe requires correct saved answers throughout the sampled visual settings in at least one Qwen3-VL size; the grounding-qualified probe additionally requires positive temporal overlap in those settings. Minimum-vision candidates must succeed in the vision-enabled settings, while text-based candidates also succeed under the black-video control. We classify these two candidate sets separately and deduplicate their question IDs, yielding 149 and 49 labeled questions, respectively. The grounded set is not necessarily a subset of the answer-only labels because classification is performed separately; these counts do not estimate intrinsic shortcut prevalence.

C.1 Shortcut Annotation Validation

We sampled 50 proposed annotations from each of the four categories across VideoMME, Video-MMMU, and LVBench. One annotator received the question, choices, labeled answer, model response, and proposed category, then flagged inappropriate assignments. Table 6 reports 192 accepted assignments out of 200. This assesses the plausibility of sampled positive labels; it does not estimate recall over undetected shortcuts or certify a model's internal reasoning.

Rejected world-knowledge labels sometimes relied on a fact mentioned in a long response without showing that it supported the answer. Some answer-choice labels instead contained hallucinated visual observations, and some superficial-recognition labels mixed knowledge with limited visual evidence. No persistent-detail assignment in the 50-item sample was rejected. These failure modes motivate using the labels as operational strata while keeping the model inputs and thinking comparisons explicit.

C.2 Additional Shortcut and Reasoning Analyses

The FPS progression analysis is retained as supporting evidence. Table 7 reports whole-benchmark accuracy across input settings. Table 4 and Appendix F.2 provide the shortcut-stratified comparisons, paired intervals, and output-validity rates.

Table 7: **Whole-benchmark accuracy across visual inputs.** Qwen3-VL accuracy (%) includes every shortcut and non-shortcut question, with invalid answers counted as wrong; NR/R denote thinking disabled/enabled, deltas are absolute points, and bold marks the better NR/R value. Endpoints uses two endpoint frames, VideoMMMU Black video retains Adaptation images, and matching LVBench 2 FPS results are unavailable.

Benchmark	Size	n	Mode	Black video	Endpoints	0.02 FPS	0.2 FPS	2 FPS
VideoMME	2B	2700	NR	27.70	**33.85**	38.63	51.48	56.22
			R	**28.89**	33.81	**38.74**	**54.30**	**60.78**
			Δ	+1.19	−0.04	+0.11	+2.81	+4.56
	4B	2700	NR	27.89	36.33	43.93	59.44	65.74
			R	**28.33**	**37.11**	**44.96**	**61.70**	**67.15**
			Δ	+0.44	+0.78	+1.04	+2.26	+1.41
	8B	2700	NR	**30.15**	**37.26**	**43.15**	**58.41**	**66.19**
			R	23.48	26.78	35.07	53.15	62.67
			Δ	−6.67	−10.48	−8.07	−5.26	−3.52
Video-MMMU	2B	900	NR	**29.44**	**31.00**	35.67	37.67	39.89
			R	27.11	27.78	**42.56**	**50.11**	**48.67**
			Δ	−2.33	−3.22	+6.89	+12.44	+8.78
	4B	900	NR	28.67	30.56	**44.00**	48.67	47.67
			R	**33.33**	**30.67**	43.89	**56.00**	**58.00**
			Δ	+4.67	+0.11	−0.11	+7.33	+10.33
	8B	900	NR	37.11	39.78	47.44	55.00	53.33
			R	**41.33**	**40.44**	**56.22**	**63.56**	**65.33**
			Δ	+4.22	+0.67	+8.78	+8.56	+12.00
LVBench	2B	1549	NR	22.53	26.66	34.80	43.83	—
			R	**23.63**	**27.18**	**37.90**	**47.00**	—
			Δ	+1.10	+0.52	+3.10	+3.16	—
	4B	1549	NR	25.18	25.50	36.80	45.77	—
			R	**26.66**	**28.53**	**41.19**	**49.19**	—
			Δ	+1.48	+3.03	+4.39	+3.42	—
	8B	1549	NR	**26.99**	**26.86**	**36.28**	**50.10**	—
			R	17.62	16.46	20.53	39.83	—
			Δ	−9.36	−10.39	−15.75	−10.26	—

C.3 TGVMME Human Validation

A single annotator examined 150 sampled TGVMME questions. The form checked whether the question was appropriate to the video, whether its options matched the question and video, and whether the labeled answer was correct. An item failed if any criterion failed; 139 passed and 11 failed, giving a sampled pass rate of 92.7%. The principal reported problem was answer ambiguity rather than a uniformly incorrect generation pattern.

For example, a question about a ladder and a wooden structure permitted both "leaning against the side" and "positioned at the base" as plausible choices. Other flagged cases involved unclear posture or contact details. These examples motivate preserving the distinction between a plausible generated question and an unambiguous single-answer test item. The evaluation uses all 1,500 fixed benchmark questions without retrospective manual exclusions from any model's denominator.

The reference boundaries come from human temporal video grounding annotations. However, the quality form did not independently redraw these boundaries or separately label whether every generated semantic detail was entailed within the designated interval. Interval-conditioned generation and whole-video item checks therefore provide complementary evidence, not an exhaustive segment-level semantic validation. The reported validation summarizes the annotator's aggregate item-quality judgments.

C.4 Inference Cost Analysis

Figure 4 reports output-length profiles across input settings. Thinking-enabled outputs can exceed 3,000 tokens under sparse input and become shorter with denser visual evidence, while no-thinking outputs remain shorter. Output length is relevant to autoregressive decoding work but does not include video decoding, visual encoding, prefill, or batching effects. It is therefore a partial cost indicator rather than a substitute for total computation.

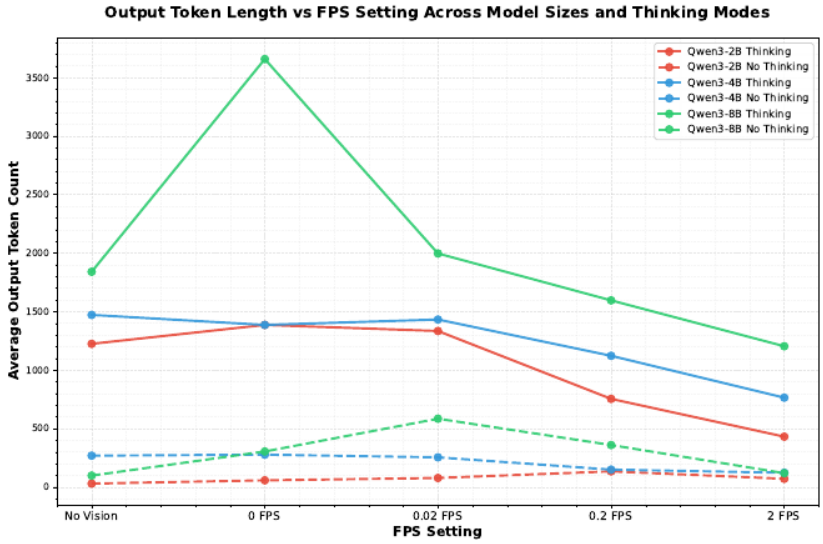


Figure 4: **Output length depends on both thinking mode and visual input.** The measured profiles show longer thinking-enabled outputs under sparse evidence; token counts describe decoding demand rather than total inference cost. The axis labels "No Vision" and "0 FPS" denote black-video and endpoint controls, respectively.

The VideoMME timing profile in Figure 5 shows that less visual input need not yield shorter inference when output generation grows. For example, the 2B thinking profile is approximately 57 seconds per question with removed video and 23 seconds at 1 FPS. This descriptive profile illustrates the observed latency pattern without normalization for hardware, batching, or realized-token budgets. Our accuracy comparison consequently makes no FLOP-normalized or latency-normalized superiority claim.

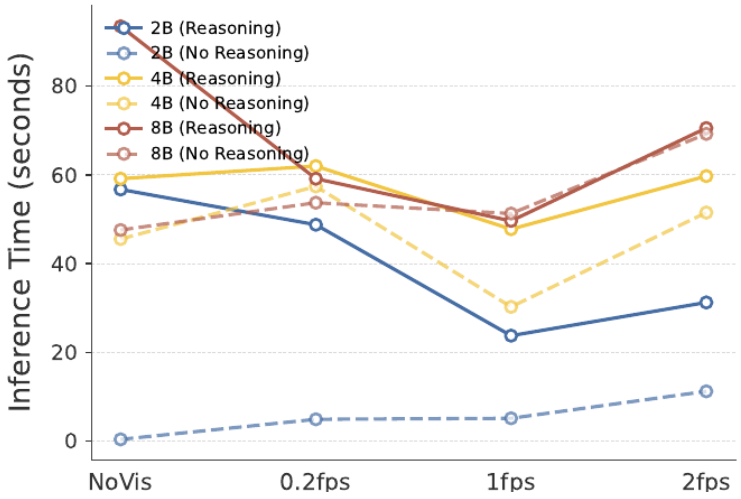


Figure 5: **Lower visual input can coincide with longer measured inference.** These VideoMME profiles include the cost of generated output and are descriptive measurements, not a controlled equal-compute comparison of input and output scaling.

C.5 Data and Code Artifacts

The benchmark data are in `artifacts/tgvmme/tgvmme.jsonl`; VLM inference scripts are in `artifacts/tgvmme/`; shortcut annotations are in `artifacts/shortcut_annotations/`. The videos are inherited from TVGBench, so the artifact README instructs readers to place the TVG-Bench videos beside the released JSONL before running inference.

C.6 Limitations and Broader Impact

Our analysis builds on open-source VLMs and relies on transparent inference and video-preprocessing pipelines. The matched comparisons concern configured thinking modes at fixed sampled frames, not all purpose-trained reasoners or adaptive visual search. Shortcut labels characterize observed solution routes, while the sampled human checks do not establish exhaustive interval-level support for generated questions. Developing diagnostic protocols that do not rely on access to upstream sampling design is an important direction for future work.

The work improves evaluation rigor for video-language models by reducing incentives to exploit textual or single-frame shortcuts. Better grounded evaluation may help identify systems that fail to localize evidence, which is important for domains where video claims require traceability. Potential negative impacts include overconfidence in benchmark scores if users ignore these limitations, or misuse of stronger video understanding systems in surveillance and other sensitive settings. We do not release new models; the released assets are benchmark data, annotations, and evaluation scripts intended for research evaluation.

D Evaluation Protocol

This appendix specifies the interventions, scoring conventions, and uncertainty estimates used in the main text. The objective is to separate changes in supplied visual evidence from the end-to-end effect of enabling thinking at a fixed visual input.

D.1 Input and Generation Controls

The existing-benchmark comparisons use Qwen3-VL-2B, 4B, and 8B thinking checkpoints in both configured modes. The no-thinking condition closes the thinking block immediately after the chat template, while the thinking condition permits generated reasoning before the answer. Gemma-4 comparisons use the template's `enable_thinking` control on E2B, E4B, and 31B checkpoints. Thus a matched pair shares model weights and input content but differs in the configured generation interface; it is not a comparison between independently post-trained models.

The black-video control replaces the source video with a 1.125-second black clip that yields two frames on the Qwen sampling path. It retains visual tokens, and Video-MMMU Adaptation questions additionally retain their static question image. The endpoint condition uses the first and last real frames because the minimum frame count is rounded to the sampler's factor of two. For positive requested FPS, the Qwen-compatible preprocessing uses uniformly spaced, endpoint-inclusive indices, a 768-frame cap, and a total-pixel budget that can lower per-frame resolution as the frame count grows. Requested FPS is consequently an intervention on the sampling budget, not a claim that every clip has an uncapped realized rate or constant spatial resolution. The TGVMME tables label the minimum-input condition as Min; its numerical metadata key is `0.002`, distinct from the four requested-FPS settings analyzed in the main text. At 2 FPS, 1,521 of 2,700 VideoMME questions and 561 of 900 Video-MMMU questions reach the 768-frame cap; at 0.2 FPS, 730 of 1,549 LVBench questions reach it.

The inference implementation defaults to a budget of 8,192 generated tokens and supports greedy or sampled decoding and explicit seeds. Its sampled-decoding defaults are temperature 0.6, top-p 0.95, top-k 20, and no fixed seed. The available outputs support one matched comparison per condition rather than repeated-seed estimates. Missing outputs and inference failures stay in the evaluation population and score as incorrect.

Frame limits and resolution adaptation control input size upstream of generation; the inference code uses these controls rather than prompt truncation. A representative Qwen3-VL-2B thinking run on Video-MMMU at 0.2 FPS records a maximum input length of 111,308 tokens, below its 128,000-token budget. Neither these controls nor the latency profiles in Appendix C.4 make the two interventions equal in computational cost. The sparse-input intervention changes both available evidence and the frame-density distribution, which this comparison evaluates together.

D.2 Answer Extraction and Temporal Scoring

For the existing-benchmark question-answering outputs, extraction prioritizes explicit answer tags, structured answer fields, recognizable malformed tags, final-answer markers, and terminal answers. Candidates must match the question's available options; conflicting or absent answers are invalid. Video-MMMU is scored using split-aware question keys, so Perception, Comprehension, and Adaptation questions remain distinct. Its option counts vary, and some questions are open-ended; a four-choice chance level is therefore not an appropriate reference for the full set.

For TGVMME, the requested response is a JSON object with `answer` and `timerange` fields. The strict parser accepts a valid structured answer or a unique malformed JSON-like quoted answer field, but not an unrestricted answer guessed from the reasoning prose. Time values may be numeric seconds or clock-formatted strings. Conflicting answers, conflicting intervals, missing fields, and reversed interval endpoints are rejected for the corresponding metric. The demonstration object, answer D with interval [12.0, 17.5], is ignored only when another structured candidate is present; a sole matching output is retained as an explicit prediction.

All matched accuracy and joint-grounding scores use the full 1,500-question denominator. The separately marked Video-R1 2 FPS absolute scores in Table 5 describe only its 128 available predictions. A wrong or invalid answer fails joint grounding even when its predicted interval

overlaps the reference. Mean tIoU and mean IoP, in contrast, average localization over all questions independently of answer correctness, with missing or invalid intervals assigned zero. For a zero-length predicted interval inside the reference interval, the scoring convention assigns IoP one and tIoU zero for a nonzero-length reference. This point-prediction convention is another reason to report tIoU alongside IoP rather than equate IoP alone with recovery of the full event.

Positive-overlap accuracy requires strictly positive intersection length, while thresholded scores require a correct answer and tIoU or IoP at least the stated threshold. We retain thresholds 0.1, 0.3, 0.5, and 0.7 in the complete tables. Out-of-bounds predictions are evaluated as emitted, without clipping them to the video duration. The alternative existing-benchmark extraction analysis rejects outputs consisting only of the prompt's demonstration object; it tests the dependence of the aggregate contrast on this specific scoring choice.

D.3 Paired Uncertainty and Aggregation

For each condition, let $d_i = y_i^{\mathrm{R}} - y_i^{\mathrm{NR}}$ be the difference in a question's binary correctness under thinking and no-thinking modes. We report $100\overline{d}$ and percentile 95% intervals from 10,000 bootstrap resamples of paired questions. The pair remains intact in each resample, and invalid outputs remain zero-valued outcomes rather than being removed. These intervals characterize question-sampling uncertainty conditional on the evaluated outputs, not variability across repeated stochastic generations.

For the aggregate shortcut contrast, we first average d_i within each question over the three Qwen sizes and four input conditions in Table 4. The difference between the shortcut-group mean and the remaining-group mean then measures a question-weighted interaction. Bootstrap sampling occurs at the question level within groups, preserving all repeated conditions for a selected question. The combined estimate weights questions equally rather than giving the three benchmarks equal weight. All per-condition and subgroup intervals are exploratory and unadjusted for multiple comparisons; intervals containing zero do not establish exact equivalence. Questions sharing a source video are not resampled as a video-level cluster, so the intervals may understate uncertainty due to shared content.

E Benchmark Construction and Annotation

TGVMME combines generated multiple-choice questions with existing human temporal annotations. This section describes what the construction and validation establish, separating interval-conditioned generation from independent human verification of generated details.

E.1 Interval-Conditioned Question Generation

The source consists of 800 TVGBench annotations: 168 from ActivityNet, 138 from Charades, 168 from EgoNLQ, 158 from HiRest, and 168 from TACoS [57]. Qwen3-VL-32B-Thinking receives the source video, retrieval sentence, exact reference interval, and video duration. The prompt requests three to four self-contained questions with four to eight options, plausible distractors, and no explicit temporal hints. Questions must concern events described by the source sentence within its interval rather than merely paraphrase the sentence.

The same generation call produces the questions, correct choices, distractors, and supporting-cue descriptions. Structural checks remove malformed multiple-choice items and unusable video references; behavioral filters use Qwen3-VL-2B and 8B outputs under altered visual inputs. These filters act on answer correctness, not independently verified temporal-necessity judgments. The final candidate pool contains 1,740 items, of which 240 are excluded to form the fixed 1,500-question set evaluated here. This finite selection does not exhaust all detected shortcut candidates, and 149 final-set questions retain at least one shortcut annotation.

Generation and filtering are model-dependent, and the final set is not intended to be an exhaustive or uniformly sampled representation of the source annotations. The supplied question set fixes the benchmark population independently of rerunning a stochastic generator. All matched TGVMME tables in this paper use that same population; construction candidates are not additional evaluation items. The human validation in Appendix C.3 provides direct, sampled evidence of question quality.

F Complete Paired Results

The following tables provide the complete matched results behind the main analysis. All TGVMME scores use the fixed 1,500-question population, with absolute no-thinking and thinking scores and all-question mean overlaps across all six input settings. Paired intervals are displayed for 0.02, 0.2, 1, and 2 FPS; the numerical files prepared for release additionally include the black-video and minimum-input intervals.

F.1 TGVMME by Model

Table 8 distinguishes joint from conditional localization in the dense settings. Tables 21 and 22 report output-validity rates, completing the interpretation of these matched comparisons.

Table 8: **Conditional grounding uses each mode's correct-answer subset.** Answer and joint IoP@0.5 use all 1,500 TGVMME questions under strict parsing; conditional IoP@0.5 uses each mode's correct answers, so its mode differences are descriptive, not paired causal localization effects. Values are percentages, means cover all questions with invalid spans scoring zero, NR/R denote thinking disabled/enabled, and bold marks the better mode within each model and FPS.

Model	FPS	Mode	Answer	Joint IoP@0.5	Conditional IoP@0.5	Mean tIoU	Mean IoP
Qwen3-VL-2B	1	NR	53.93	28.73	53.28	27.74	51.02
	1	R	**55.80**	**32.00**	**57.35**	**28.15**	**52.51**
	2	NR	55.87	32.20	57.64	29.02	53.41
	2	R	**58.47**	**35.27**	**60.32**	**30.43**	**55.66**
Qwen3-VL-4B	1	NR	61.60	38.73	62.88	**30.87**	**58.94**
	1	R	**63.93**	**40.47**	**63.30**	29.17	58.66
	2	NR	63.80	41.93	65.73	32.29	61.39
	2	R	**68.20**	**45.73**	**67.06**	**32.35**	**61.84**
Qwen3-VL-8B	1	NR	64.67	40.93	63.30	**32.47**	**59.84**
	1	R	**64.80**	**41.53**	**64.09**	28.52	55.24
	2	NR	66.80	44.07	**65.97**	**32.60**	**61.78**
	2	R	**67.27**	**44.20**	65.71	29.45	57.42
Gemma-4-E2B	1	NR	37.53	11.33	30.20	6.45	30.51
	1	R	**45.67**	**14.60**	**31.97**	**10.31**	**31.28**
	2	NR	36.93	11.07	29.96	4.70	28.42
	2	R	**44.20**	**13.80**	**31.22**	**10.23**	**30.51**
Gemma-4-E4B	1	NR	49.73	16.67	33.51	6.67	33.36
	1	R	**56.00**	**23.67**	**42.26**	**17.39**	**41.45**
	2	NR	49.87	17.00	34.09	6.21	33.11
	2	R	**53.47**	**23.80**	**44.51**	**18.39**	**41.61**
Gemma-4-31B	1	NR	**70.67**	**43.47**	**61.51**	**26.47**	**58.01**
	1	R	70.27	40.53	57.69	24.42	54.47
	2	NR	70.00	**41.93**	**59.90**	**25.51**	**55.72**
	2	R	**70.53**	41.87	59.36	24.65	54.23

An in-target point prediction receives IoP $= 1$ but tIoU $= 0$ for a positive-length reference; IoP alone does not measure event coverage.

Tables 9 and 10 report Qwen3-VL-2B absolute scores and paired effects.

Table 9: **Complete answer and grounding scores for Qwen3-VL-2B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	21.73	6.93	5.27	2.13	0.67	0.27
Black	R	23.27	7.07	5.40	3.20	1.60	0.53
Min	NR	24.00	7.73	5.13	2.73	1.40	0.87
Min	R	21.73	7.00	5.07	2.60	1.33	0.60
0.02	NR	24.60	8.07	5.13	2.67	1.27	0.40
0.02	R	21.40	6.87	5.27	2.73	1.47	0.80
0.2	NR	37.67	16.60	13.27	7.73	3.53	1.20
0.2	R	38.53	18.47	15.60	7.47	3.40	1.47
1	NR	53.93	39.67	35.00	21.60	11.27	5.13
1	R	55.80	44.00	38.93	24.47	13.20	6.27
2	NR	55.87	42.60	37.07	22.27	13.67	7.20
2	R	**58.47**	**48.07**	**42.20**	**25.60**	**16.27**	**8.53**

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	6.73	6.20	5.53	5.33	8.07	25.25
Black	R	6.53	5.80	5.73	4.80	8.38	22.73
Min	NR	6.87	4.87	4.60	4.27	8.97	21.47
Min	R	6.53	5.13	4.73	4.40	7.62	18.33
0.02	NR	7.07	5.13	4.73	4.00	8.77	21.75
0.02	R	6.27	5.27	4.67	3.93	7.84	19.43
0.2	NR	15.53	13.07	11.67	10.27	13.44	30.30
0.2	R	17.27	14.20	12.20	10.00	13.01	28.54
1	NR	37.93	33.33	28.73	23.33	27.74	51.02
1	R	42.00	36.47	32.00	27.60	28.15	52.51
2	NR	40.80	35.60	32.20	27.13	29.02	53.41
2	R	**45.87**	**40.20**	**35.27**	**30.60**	**30.43**	**55.66**

Table 10: **Paired reasoning effects for Qwen3-VL-2B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	-3.20	$[-5.73, -0.67]$	$+0.87$	$[-1.60, +3.47]$	$+1.87$	$[-0.47, +4.27]$	$\mathbf{+2.60}$	$[+0.40, +4.87]$
Positive overlap	-1.20	$[-2.87, +0.40]$	$+1.87$	$[-0.20, +3.87]$	$+4.33$	$[+1.87, +6.80]$	$\mathbf{+5.47}$	$[+3.13, +7.80]$
tIoU@0.1	$+0.13$	$[-1.20, +1.47]$	$+2.33$	$[+0.40, +4.27]$	$+3.93$	$[+1.33, +6.47]$	$\mathbf{+5.13}$	$[+2.80, +7.53]$
tIoU@0.3	$+0.07$	$[-0.87, +1.07]$	-0.27	$[-1.80, +1.33]$	$+2.87$	$[+0.60, +5.07]$	$\mathbf{+3.33}$	$[+1.33, +5.33]$
tIoU@0.5	$+0.20$	$[-0.40, +0.87]$	-0.13	$[-1.27, +1.00]$	$+1.93$	$[+0.13, +3.73]$	$\mathbf{+2.60}$	$[+0.87, +4.33]$
tIoU@0.7	$+0.40$	$[+0.00, +0.87]$	$+0.27$	$[-0.47, +1.00]$	$+1.13$	$[-0.13, +2.40]$	$\mathbf{+1.33}$	$[+0.00, +2.60]$
IoP@0.1	-0.80	$[-2.33, +0.67]$	$+1.73$	$[-0.27, +3.67]$	$+4.07$	$[+1.60, +6.53]$	$\mathbf{+5.07}$	$[+2.73, +7.40]$
IoP@0.3	$+0.13$	$[-1.20, +1.47]$	$+1.13$	$[-0.80, +3.00]$	$+3.13$	$[+0.73, +5.53]$	$\mathbf{+4.60}$	$[+2.33, +6.87]$
IoP@0.5	-0.07	$[-1.33, +1.20]$	$+0.53$	$[-1.27, +2.27]$	$\mathbf{+3.27}$	$[+1.07, +5.53]$	$+3.07$	$[+0.93, +5.27]$
IoP@0.7	-0.07	$[-1.27, +1.13]$	-0.27	$[-1.93, +1.33]$	$\mathbf{+4.27}$	$[+2.13, +6.40]$	$+3.47$	$[+1.47, +5.40]$

Tables 11 and 12 report Qwen3-VL-4B absolute scores and paired effects.

Table 11: **Complete answer and grounding scores for Qwen3-VL-4B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	24.60	8.20	6.07	3.00	1.27	0.27
Black	R	25.07	8.27	5.93	3.20	1.27	0.33
Min	NR	28.00	11.33	9.20	4.53	2.73	1.33
Min	R	29.00	9.20	5.93	2.93	1.33	0.53
0.02	NR	27.67	10.87	7.93	4.73	2.07	0.87
0.02	R	28.33	8.73	6.13	2.93	1.20	0.60
0.2	NR	43.87	23.13	19.20	9.80	4.93	2.20
0.2	R	44.27	23.40	17.67	9.00	4.60	1.73
1	NR	61.60	49.40	43.67	27.80	16.80	8.00
1	R	63.93	51.07	45.40	28.53	16.27	6.93
2	NR	63.80	53.07	47.20	30.67	17.20	8.87
2	R	**68.20**	**57.07**	**51.73**	**33.20**	**20.40**	**9.47**

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	7.60	6.53	6.27	6.00	8.61	25.68
Black	R	7.53	6.67	6.33	6.13	8.26	24.29
Min	NR	10.47	8.67	7.67	6.40	10.84	24.50
Min	R	8.67	7.13	6.53	5.73	7.67	22.47
0.02	NR	9.67	7.87	6.53	5.73	10.68	24.10
0.02	R	8.07	6.53	5.67	4.67	7.70	22.12
0.2	NR	21.67	18.07	15.53	12.67	16.24	35.77
0.2	R	22.20	18.60	15.67	13.87	14.54	35.39
1	NR	47.53	42.13	38.73	34.00	30.87	58.94
1	R	49.20	44.33	40.47	35.33	29.17	58.66
2	NR	51.53	46.47	41.93	36.47	32.29	61.39
2	R	**55.73**	**50.27**	**45.73**	**39.80**	**32.35**	**61.84**

Table 12: **Paired reasoning effects for Qwen3-VL-4B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	$+0.67$	$[-1.53, +2.93]$	$+0.40$	$[-2.07, +2.87]$	$+2.33$	$[+0.13, +4.67]$	$\mathbf{+4.40}$	$[+2.27, +6.60]$
Positive overlap	-2.13	$[-3.73, -0.53]$	$+0.27$	$[-2.00, +2.53]$	$+1.67$	$[-0.67, +4.00]$	$\mathbf{+4.00}$	$[+1.80, +6.13]$
tIoU@0.1	-1.80	$[-3.20, -0.47]$	-1.53	$[-3.67, +0.53]$	$+1.73$	$[-0.67, +4.13]$	$\mathbf{+4.53}$	$[+2.27, +6.73]$
tIoU@0.3	-1.80	$[-2.93, -0.73]$	-0.80	$[-2.60, +0.93]$	$+0.73$	$[-1.33, +2.87]$	$\mathbf{+2.53}$	$[+0.47, +4.60]$
tIoU@0.5	-0.87	$[-1.67, -0.07]$	-0.33	$[-1.67, +1.00]$	-0.53	$[-2.40, +1.33]$	$\mathbf{+3.20}$	$[+1.33, +5.07]$
tIoU@0.7	-0.27	$[-0.80, +0.27]$	-0.47	$[-1.27, +0.33]$	-1.07	$[-2.47, +0.33]$	$\mathbf{+0.60}$	$[-0.80, +1.93]$
IoP@0.1	-1.60	$[-3.07, -0.13]$	$+0.53$	$[-1.73, +2.80]$	$+1.67$	$[-0.73, +4.07]$	$\mathbf{+4.20}$	$[+2.00, +6.40]$
IoP@0.3	-1.33	$[-2.73, +0.00]$	$+0.53$	$[-1.53, +2.60]$	$+2.20$	$[+0.00, +4.47]$	$\mathbf{+3.80}$	$[+1.67, +5.93]$
IoP@0.5	-0.87	$[-2.13, +0.40]$	$+0.13$	$[-1.80, +2.07]$	$+1.73$	$[-0.60, +4.00]$	$\mathbf{+3.80}$	$[+1.67, +5.93]$
IoP@0.7	-1.07	$[-2.33, +0.20]$	$+1.20$	$[-0.60, +3.00]$	$+1.33$	$[-0.80, +3.47]$	$\mathbf{+3.33}$	$[+1.20, +5.40]$

Tables 13 and 14 report Qwen3-VL-8B absolute scores and paired effects.

Table 13: **Complete answer and grounding scores for Qwen3-VL-8B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	26.67	8.53	6.13	2.33	0.87	0.07
Black	R	25.27	7.93	4.53	1.40	0.67	0.13
Min	NR	27.73	9.27	7.00	3.13	1.47	0.67
Min	R	23.73	6.93	4.67	2.13	1.20	0.60
0.02	NR	28.60	10.33	7.73	3.73	2.07	0.80
0.02	R	24.13	7.33	4.53	2.27	1.00	0.47
0.2	NR	43.87	25.73	22.40	12.27	5.87	2.13
0.2	R	42.20	25.47	21.20	10.67	4.93	1.87
1	NR	64.67	52.80	48.20	31.33	18.47	8.80
1	R	64.80	54.67	49.07	30.20	16.27	7.93
2	NR	66.80	55.27	49.67	**32.27**	**18.73**	**9.60**
2	R	**67.27**	**57.47**	**51.07**	31.00	16.93	9.07

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	8.07	7.20	6.27	5.93	8.37	27.79
Black	R	7.60	7.00	6.60	6.27	4.48	20.83
Min	NR	8.67	7.20	6.47	5.67	9.08	24.89
Min	R	7.87	6.60	6.00	5.33	5.05	18.53
0.02	NR	10.00	8.07	7.67	6.47	8.77	24.50
0.02	R	7.47	6.47	5.67	4.87	5.17	20.14
0.2	NR	24.40	20.67	18.27	15.87	18.44	38.87
0.2	R	23.80	20.20	17.93	15.13	14.33	32.70
1	NR	51.33	45.40	40.93	34.73	32.47	59.84
1	R	52.40	46.87	41.53	35.87	28.52	55.24
2	NR	54.07	48.33	44.07	39.00	**32.60**	**61.78**
2	R	**55.80**	**49.00**	**44.20**	**39.07**	29.45	57.42

Table 14: **Paired reasoning effects for Qwen3-VL-8B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	-4.47	$[-6.93, -2.00]$	-1.67	$[-4.13, +0.87]$	$+0.13$	$[-2.20, +2.47]$	$\mathbf{+0.47}$	$[-1.73, +2.67]$
Positive overlap	-3.00	$[-4.73, -1.33]$	-0.27	$[-2.47, +1.87]$	$+1.87$	$[-0.47, +4.20]$	$\mathbf{+2.20}$	$[-0.07, +4.40]$
tIoU@0.1	-3.20	$[-4.67, -1.80]$	-1.20	$[-3.33, +1.00]$	$+0.87$	$[-1.67, +3.33]$	$\mathbf{+1.40}$	$[-0.87, +3.73]$
tIoU@0.3	-1.47	$[-2.60, -0.33]$	-1.60	$[-3.47, +0.27]$	$\mathbf{-1.13}$	$[-3.40, +1.13]$	-1.27	$[-3.40, +0.87]$
tIoU@0.5	-1.07	$[-1.87, -0.27]$	$\mathbf{-0.93}$	$[-2.27, +0.47]$	-2.20	$[-4.13, -0.20]$	-1.80	$[-3.67, +0.13]$
tIoU@0.7	-0.33	$[-0.87, +0.13]$	$\mathbf{-0.27}$	$[-1.07, +0.53]$	-0.87	$[-2.40, +0.67]$	-0.53	$[-2.00, +1.00]$
IoP@0.1	-2.53	$[-4.20, -0.93]$	-0.60	$[-2.80, +1.53]$	$+1.07$	$[-1.33, +3.53]$	$\mathbf{+1.73}$	$[-0.67, +4.00]$
IoP@0.3	-1.60	$[-3.13, -0.07]$	-0.47	$[-2.53, +1.60]$	$\mathbf{+1.47}$	$[-0.93, +3.80]$	$+0.67$	$[-1.60, +2.93]$
IoP@0.5	-2.00	$[-3.47, -0.53]$	-0.33	$[-2.27, +1.60]$	$\mathbf{+0.60}$	$[-1.80, +2.93]$	$+0.13$	$[-2.07, +2.40]$
IoP@0.7	-1.60	$[-2.93, -0.27]$	-0.73	$[-2.53, +1.00]$	$\mathbf{+1.13}$	$[-1.00, +3.27]$	$+0.07$	$[-2.00, +2.20]$

Tables 15 and 16 report Gemma-4-E2B absolute scores and paired effects.

Table 15: **Complete answer and grounding scores for Gemma-4-E2B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	19.87	4.73	2.47	0.33	0.00	0.00
Black	R	22.60	5.27	2.93	0.40	0.00	0.00
Min	NR	26.00	11.67	8.53	3.80	0.80	0.27
Min	R	28.07	15.33	11.80	4.87	0.93	0.27
0.02	NR	25.07	11.40	8.53	3.87	1.13	0.20
0.02	R	26.87	14.00	10.47	4.40	0.93	0.20
0.2	NR	35.53	10.13	4.27	2.00	0.80	0.47
0.2	R	40.13	15.07	10.47	6.00	1.80	0.80
1	NR	37.53	13.07	5.67	3.40	1.87	0.73
1	R	**45.67**	**20.53**	**11.93**	6.13	**2.67**	**1.27**
2	NR	36.93	10.53	4.73	2.33	1.27	0.67
2	R	44.20	19.67	11.47	**6.20**	2.27	1.07

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	4.60	4.53	4.20	4.20	3.60	24.49
Black	R	5.13	5.07	4.87	4.87	3.73	24.54
Min	NR	9.87	6.33	3.67	3.20	10.15	20.97
Min	R	13.87	8.60	4.87	4.53	11.72	23.02
0.02	NR	9.87	6.87	4.47	3.67	10.26	21.59
0.02	R	12.40	8.00	4.87	4.20	**12.08**	23.91
0.2	NR	9.93	9.27	8.13	7.73	4.97	24.55
0.2	R	15.07	12.67	9.13	8.27	10.23	25.12
1	NR	13.47	12.33	11.33	10.87	6.45	30.51
1	R	**19.80**	**17.40**	**14.60**	**13.07**	10.31	**31.28**
2	NR	12.73	11.80	11.07	10.87	4.70	28.42
2	R	18.73	16.33	13.80	12.67	10.23	30.51

Table 16: **Paired reasoning effects for Gemma-4-E2B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	$+1.80$	$[-0.53, +4.20]$	$+4.60$	$[+2.07, +7.27]$	$\mathbf{+8.13}$	$[+5.40, +10.93]$	$+7.27$	$[+4.47, +10.07]$
Positive overlap	$+2.60$	$[+0.60, +4.53]$	$+4.93$	$[+3.00, +6.87]$	$+7.47$	$[+5.27, +9.60]$	$\mathbf{+9.13}$	$[+6.93, +11.33]$
tIoU@0.1	$+1.93$	$[+0.13, +3.67]$	$+6.20$	$[+4.53, +8.00]$	$+6.27$	$[+4.40, +8.13]$	$\mathbf{+6.73}$	$[+4.93, +8.60]$
tIoU@0.3	$+0.53$	$[-0.73, +1.80]$	$\mathbf{+4.00}$	$[+2.73, +5.33]$	$+2.73$	$[+1.33, +4.20]$	$+3.87$	$[+2.53, +5.27]$
tIoU@0.5	-0.20	$[-0.87, +0.47]$	$\mathbf{+1.00}$	$[+0.27, +1.80]$	$+0.80$	$[-0.20, +1.80]$	$\mathbf{+1.00}$	$[+0.07, +1.93]$
tIoU@0.7	$+0.00$	$[-0.27, +0.27]$	$+0.33$	$[-0.13, +0.87]$	$\mathbf{+0.53}$	$[-0.13, +1.20]$	$+0.40$	$[-0.27, +1.07]$
IoP@0.1	$+2.53$	$[+0.73, +4.40]$	$+5.13$	$[+3.20, +7.00]$	$\mathbf{+6.33}$	$[+4.27, +8.40]$	$+6.00$	$[+3.93, +8.07]$
IoP@0.3	$+1.13$	$[-0.40, +2.73]$	$+3.40$	$[+1.60, +5.20]$	$\mathbf{+5.07}$	$[+3.00, +7.13]$	$+4.53$	$[+2.53, +6.47]$
IoP@0.5	$+0.40$	$[-0.87, +1.67]$	$+1.00$	$[-0.53, +2.60]$	$\mathbf{+3.27}$	$[+1.40, +5.13]$	$+2.73$	$[+0.93, +4.53]$
IoP@0.7	$+0.53$	$[-0.60, +1.67]$	$+0.53$	$[-0.93, +2.00]$	$\mathbf{+2.20}$	$[+0.47, +4.00]$	$+1.80$	$[+0.07, +3.53]$

Tables 17 and 18 report Gemma-4-E4B absolute scores and paired effects.

Table 17: **Complete answer and grounding scores for Gemma-4-E4B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	18.60	1.60	0.73	0.20	0.07	0.00
Black	R	22.73	4.80	2.80	0.67	0.13	0.00
Min	NR	28.40	11.07	7.93	4.33	1.80	0.53
Min	R	30.40	11.00	8.27	5.20	1.93	0.53
0.02	NR	28.40	10.40	7.60	5.07	2.27	0.80
0.02	R	32.00	13.13	10.47	5.47	2.00	0.93
0.2	NR	43.60	15.07	8.73	5.40	2.60	1.13
0.2	R	45.80	22.27	17.93	11.40	4.67	2.33
1	NR	49.73	17.20	9.20	4.20	2.07	0.73
1	R	**56.00**	**32.53**	**26.27**	**15.00**	7.20	2.87
2	NR	49.87	16.20	8.27	3.93	2.00	0.73
2	R	53.47	30.13	24.93	14.80	**7.93**	**3.67**

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	4.33	4.13	3.87	3.87	1.58	24.05
Black	R	4.93	4.87	4.40	4.33	4.29	24.51
Min	NR	10.20	7.93	5.93	5.27	9.66	23.83
Min	R	9.53	7.13	4.67	3.53	10.87	21.53
0.02	NR	9.80	8.47	6.33	5.67	10.14	23.90
0.02	R	12.27	8.73	5.87	4.53	10.22	21.00
0.2	NR	16.27	14.67	12.40	11.40	8.34	28.90
0.2	R	21.00	18.00	14.00	12.07	14.86	30.57
1	NR	18.40	17.93	16.67	16.07	6.67	33.36
1	R	**30.93**	**27.53**	23.67	21.20	17.39	41.45
2	NR	18.07	17.87	17.00	16.60	6.21	33.11
2	R	29.67	27.20	**23.80**	**21.33**	**18.39**	**41.61**

Table 18: **Paired reasoning effects for Gemma-4-E4B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	$+3.60$	$[+1.20, +6.07]$	$+2.20$	$[-0.27, +4.67]$	$\mathbf{+6.27}$	$[+3.73, +8.87]$	$+3.60$	$[+1.13, +6.13]$
Positive overlap	$+2.73$	$[+0.73, +4.67]$	$+7.20$	$[+4.80, +9.53]$	$\mathbf{+15.33}$	$[+12.73, +17.80]$	$+13.93$	$[+11.47, +16.33]$
tIoU@0.1	$+2.87$	$[+1.07, +4.67]$	$+9.20$	$[+6.87, +11.47]$	$\mathbf{+17.07}$	$[+14.60, +19.47]$	$+16.67$	$[+14.27, +19.07]$
tIoU@0.3	$+0.40$	$[-1.07, +1.87]$	$+6.00$	$[+4.20, +7.87]$	$+10.80$	$[+8.93, +12.67]$	$\mathbf{+10.87}$	$[+8.93, +12.80]$
tIoU@0.5	-0.27	$[-1.20, +0.67]$	$+2.07$	$[+0.80, +3.40]$	$+5.13$	$[+3.73, +6.53]$	$\mathbf{+5.93}$	$[+4.47, +7.40]$
tIoU@0.7	$+0.13$	$[-0.47, +0.80]$	$+1.20$	$[+0.27, +2.07]$	$+2.13$	$[+1.27, +3.07]$	$\mathbf{+2.93}$	$[+1.93, +4.00]$
IoP@0.1	$+2.47$	$[+0.60, +4.33]$	$+4.73$	$[+2.40, +7.07]$	$\mathbf{+12.53}$	$[+10.00, +15.07]$	$+11.60$	$[+9.20, +14.07]$
IoP@0.3	$+0.27$	$[-1.40, +1.93]$	$+3.33$	$[+1.13, +5.53]$	$\mathbf{+9.60}$	$[+7.13, +12.07]$	$+9.33$	$[+6.93, +11.73]$
IoP@0.5	-0.47	$[-1.93, +1.00]$	$+1.60$	$[-0.40, +3.60]$	$\mathbf{+7.00}$	$[+4.60, +9.40]$	$+6.80$	$[+4.53, +9.07]$
IoP@0.7	-1.13	$[-2.47, +0.27]$	$+0.67$	$[-1.13, +2.47]$	$\mathbf{+5.13}$	$[+2.87, +7.40]$	$+4.73$	$[+2.60, +6.87]$

Tables 19 and 20 report Gemma-4-31B absolute scores and paired effects.

Table 19: **Complete answer and grounding scores for Gemma-4-31B.** Grounded percentages require a correct answer under strict parsing; mean overlaps ($\times 100$) include all 1,500 TGVMME questions, assigning zero to invalid spans. NR/R denote thinking disabled/enabled; Black and Min denote black-video and minimum-input controls; bold marks each metric's maximum, and in-target point predictions receive IoP $= 1$.

Input (FPS)	Mode	MCQ	Positive overlap	tIoU@0.1	tIoU@0.3	tIoU@0.5	tIoU@0.7
Black	NR	37.27	8.33	5.80	1.13	0.60	0.33
Black	R	41.13	10.80	7.80	2.73	0.93	0.20
Min	NR	44.53	21.07	17.47	9.73	4.20	2.20
Min	R	46.07	20.40	17.13	9.73	4.80	1.33
0.02	NR	46.07	21.67	17.60	10.40	4.27	1.93
0.02	R	46.93	19.47	16.73	10.47	5.33	2.07
0.2	NR	60.47	36.47	32.40	20.80	10.20	4.47
0.2	R	61.20	34.67	30.27	18.27	9.20	4.20
1	NR	**70.67**	**52.47**	**45.00**	26.13	**15.07**	**7.27**
1	R	70.27	50.80	42.73	24.60	13.67	5.87
2	NR	70.00	50.40	43.67	25.53	14.73	7.20
2	R	70.53	51.40	43.47	**26.40**	14.07	6.87

Input (FPS)	Mode	IoP@0.1	IoP@0.3	IoP@0.5	IoP@0.7	Mean tIoU	Mean IoP
Black	NR	8.07	8.00	7.93	7.67	5.09	24.93
Black	R	10.53	10.20	9.67	8.93	6.68	25.87
Min	NR	20.13	16.47	13.80	12.33	14.62	31.46
Min	R	19.33	16.00	13.60	11.53	13.76	29.84
0.02	NR	20.80	17.60	14.07	12.53	14.26	32.19
0.02	R	18.27	15.93	13.53	11.60	13.67	28.58
0.2	NR	34.87	31.53	25.13	20.80	22.50	40.90
0.2	R	32.93	28.13	22.93	18.73	19.49	35.68
1	NR	**51.73**	**47.80**	**43.47**	**39.60**	**26.47**	**58.01**
1	R	49.80	45.00	40.53	36.93	24.42	54.47
2	NR	50.00	45.67	41.93	37.87	25.51	55.72
2	R	50.27	45.87	41.87	36.87	24.65	54.23

Table 20: **Paired reasoning effects for Gemma-4-31B.** Thinking minus no-thinking deltas and paired 95% bootstrap intervals are in absolute percentage points on the same 1,500 questions under strict parsing. Intervals are unadjusted for multiple comparisons; bold marks each metric's largest displayed delta.

Metric	0.02 FPS		0.2 FPS		1 FPS		2 FPS	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
MCQ	$+\mathbf{0.87}$	$[-1.20, +3.00]$	$+0.73$	$[-1.13, +2.60]$	-0.40	$[-2.13, +1.33]$	$+0.53$	$[-1.27, +2.27]$
Positive overlap	-2.20	$[-4.27, -0.20]$	-1.80	$[-3.87, +0.27]$	-1.67	$[-3.80, +0.47]$	$+\mathbf{1.00}$	$[-1.13, +3.13]$
tIoU@0.1	-0.87	$[-2.73, +1.00]$	-2.13	$[-4.20, -0.13]$	-2.27	$[-4.47, -0.07]$	$-\mathbf{0.20}$	$[-2.40, +2.00]$
tIoU@0.3	$+0.07$	$[-1.47, +1.53]$	-2.53	$[-4.40, -0.60]$	-1.53	$[-3.60, +0.53]$	$+\mathbf{0.87}$	$[-1.20, +3.00]$
tIoU@0.5	$+\mathbf{1.07}$	$[+0.00, +2.20]$	-1.00	$[-2.53, +0.53]$	-1.40	$[-3.13, +0.33]$	-0.67	$[-2.27, +1.00]$
tIoU@0.7	$+\mathbf{0.13}$	$[-0.73, +1.00]$	-0.27	$[-1.33, +0.80]$	-1.40	$[-2.67, -0.20]$	-0.33	$[-1.60, +0.93]$
IoP@0.1	-2.53	$[-4.47, -0.53]$	-1.93	$[-3.93, +0.13]$	-1.93	$[-4.07, +0.20]$	$+\mathbf{0.27}$	$[-1.93, +2.47]$
IoP@0.3	-1.67	$[-3.53, +0.20]$	-3.40	$[-5.33, -1.47]$	-2.80	$[-5.00, -0.60]$	$+\mathbf{0.20}$	$[-2.00, +2.40]$
IoP@0.5	-0.53	$[-2.27, +1.27]$	-2.20	$[-4.13, -0.27]$	-2.93	$[-5.00, -0.80]$	$-\mathbf{0.07}$	$[-2.27, +2.20]$
IoP@0.7	$-\mathbf{0.93}$	$[-2.60, +0.73]$	-2.07	$[-3.80, -0.33]$	-2.67	$[-4.73, -0.60]$	-1.00	$[-3.13, +1.13]$

Table 21: **Invalid-output rates for Qwen3-VL.** Strict-parser invalid-answer and invalid-span rates (%) use all 1,500 TGVMME questions at each input setting, with NR/R denoting thinking disabled/enabled. Bold marks the lower rate in each NR/R pair; Black and Min denote black-video and minimum-input controls.

Model	Input	Invalid answer		Invalid span	
		NR	R	NR	R
Qwen3-VL-2B	Black	**0.00**	5.00	**0.07**	4.93
	Min	**0.00**	9.20	**0.07**	9.07
	0.02	**0.07**	9.13	**0.13**	9.20
	0.2	**0.07**	8.47	**0.27**	8.47
	1	**0.13**	2.67	**1.20**	2.67
	2	**0.27**	1.13	**1.13**	**1.13**
Qwen3-VL-4B	Black	**0.07**	1.47	**0.27**	1.73
	Min	**0.07**	0.53	**0.00**	0.67
	0.02	**0.07**	0.73	**0.33**	0.87
	0.2	**0.07**	1.67	**0.13**	1.60
	1	**0.00**	0.93	**0.00**	0.93
	2	**0.00**	0.80	**0.00**	0.80
Qwen3-VL-8B	Black	**0.07**	18.80	**0.07**	18.80
	Min	**0.00**	18.67	**0.00**	18.67
	0.02	**0.00**	18.40	**0.13**	18.40
	0.2	**0.00**	16.13	**0.00**	16.13
	1	**0.00**	8.60	**0.13**	8.60
	2	**0.00**	7.40	**0.00**	7.40

Table 22: **Invalid-output rates for Gemma-4.** Strict-parser invalid-answer and invalid-span rates (%) use all 1,500 TGVMME questions at each input setting, with NR/R denoting thinking disabled/enabled. Bold marks the lower rate in each NR/R pair; Black and Min denote black-video and minimum-input controls.

Model	Input	Invalid answer		Invalid span	
		NR	R	NR	R
Gemma-4-E2B	Black	**0.00**	3.53	**0.00**	**0.00**
	Min	**0.27**	0.87	**0.60**	0.67
	0.02	**0.20**	0.60	0.93	**0.67**
	0.2	1.40	**0.00**	**0.40**	0.47
	1	1.80	**0.13**	**0.07**	0.13
	2	1.67	**0.13**	**0.20**	**0.20**
Gemma-4-E4B	Black	**9.00**	10.60	**0.00**	1.00
	Min	0.93	**0.87**	**0.00**	0.40
	0.02	1.53	**0.40**	**0.07**	0.33
	0.2	0.07	**0.00**	**0.07**	0.27
	1	0.07	**0.00**	0.73	**0.20**
	2	**0.00**	**0.00**	1.13	**0.13**
Gemma-4-31B	Black	**0.00**	0.07	**0.00**	0.07
	Min	**0.00**	0.67	**0.00**	0.67
	0.02	**0.00**	0.40	**0.00**	0.40
	0.2	**0.00**	0.40	**0.00**	0.47
	1	**0.07**	0.67	**0.40**	0.80
	2	**0.07**	0.47	1.33	**0.53**

F.2 Fixed Shortcut Groups and Output Validity

Tables 23, 25, and 27 supplement Table 4 with paired intervals. Tables 24, 26, and 28 retain the output-validity information needed to interpret those differences. The same question groups are used in every condition, and failures do not change the denominators.

High invalid-answer rates can account for part of a negative thinking effect under a particular interface. The tables therefore establish differences in credited end-to-end performance, not a decomposition into hidden reasoning quality, formatting reliability, and sampling shift. The aggregate result below summarizes that measured behavior while retaining its benchmark-specific limits.

Table 23: **Paired shortcut and non-shortcut effects on VideoMME.** Qwen3-VL reasoning deltas and paired 95% bootstrap intervals (absolute points) use 899 shortcut (SC) and 1801 non-shortcut questions, with invalid answers counted as wrong. Intervals are unadjusted for multiple comparisons; bold marks the largest delta within each size and question group.

Size	Input	SC Δ	95% CI	Non-SC Δ	95% CI
2B	Black video	+1.78	[−2.11, +5.67]	**+0.89**	[−1.55, +3.33]
	Endpoints	**+9.23**	[+5.34, +13.13]	−4.66	[−6.83, −2.44]
	0.02 FPS	+2.78	[−0.67, +6.23]	−1.22	[−3.55, +1.11]
	0.2 FPS	+7.45	[+4.89, +10.01]	+0.50	[−1.78, +2.83]
4B	Black video	+2.89	[+0.11, +5.67]	−0.78	[−2.22, +0.72]
	Endpoints	**+11.90**	[+8.90, +15.02]	−4.78	[−6.55, −3.00]
	0.02 FPS	+3.45	[+0.67, +6.23]	−0.17	[−2.33, +2.00]
	0.2 FPS	+6.67	[+4.34, +9.01]	**+0.06**	[−2.22, +2.39]
8B	Black video	−10.23	[−13.57, −7.01]	−4.89	[−6.77, −3.00]
	Endpoints	−10.68	[−14.24, −7.12]	−10.38	[−12.33, −8.44]
	0.02 FPS	−9.34	[−12.46, −6.23]	−7.44	[−9.55, −5.27]
	0.2 FPS	**−6.34**	[−9.23, −3.45]	**−4.72**	[−7.00, −2.39]

Table 24: **Invalid-answer rates on VideoMME.** Qwen3-VL invalid-answer rates (%) are shown without (NR) and with reasoning (R), using 899 shortcut (SC) and 1801 non-shortcut questions in every condition. Bold marks the lower rate in each NR/R pair.

Size	Input	SC		Non-SC	
		NR	R	NR	R
2B	Black video	**0.11**	9.01	**0.28**	11.16
	Endpoints	**0.22**	9.01	**0.06**	14.77
	0.02 FPS	**0.00**	9.79	**0.11**	17.16
	0.2 FPS	**0.56**	2.56	**0.83**	7.55
4B	Black video	**0.22**	4.23	**0.72**	4.78
	Endpoints	**0.67**	1.78	**1.17**	4.50
	0.02 FPS	**0.22**	3.67	**0.39**	8.83
	0.2 FPS	**1.11**	1.67	**0.89**	7.88
8B	Black video	**1.00**	28.36	**1.05**	32.76
	Endpoints	**3.11**	25.03	**5.39**	40.53
	0.02 FPS	**5.45**	21.58	**11.16**	39.70
	0.2 FPS	**4.12**	12.68	**6.50**	30.65

Table 25: **Paired shortcut and non-shortcut effects on Video-MMMU.** Qwen3-VL reasoning deltas and paired 95% bootstrap intervals (absolute points) use 369 shortcut (SC) and 531 non-shortcut questions, with invalid answers counted as wrong. Intervals are unadjusted for multiple comparisons; bold marks the largest delta within each size and question group.

Size	Input	SC Δ	95% CI	Non-SC Δ	95% CI
2B	Black video	+11.11	[+5.15, +17.34]	−11.68	[−15.63, −7.72]
	Endpoints	+11.92	[+5.96, +17.89]	−13.75	[−17.70, −9.79]
	0.02 FPS	+23.04	[+17.07, +29.00]	−4.33	[−8.66, +0.19]
	0.2 FPS	**+26.56**	[+21.14, +31.98]	**+2.64**	[−1.88, +7.16]
4B	Black video	+10.57	[+5.96, +15.18]	+0.56	[−2.64, +3.77]
	Endpoints	+5.15	[+0.27, +10.03]	−3.39	[−6.21, −0.75]
	0.02 FPS	+4.07	[−0.81, +8.94]	−3.01	[−6.40, +0.38]
	0.2 FPS	**+13.82**	[+8.94, +18.70]	**+2.82**	[−1.13, +6.78]
8B	Black video	+13.28	[+7.86, +18.70]	−2.07	[−6.21, +2.07]
	Endpoints	**+20.33**	[+15.18, +25.47]	−12.99	[−17.14, −8.85]
	0.02 FPS	+15.45	[+10.84, +19.78]	**+4.14**	[−0.38, +8.66]
	0.2 FPS	+15.45	[+10.84, +19.78]	+3.77	[−0.75, +8.29]

Table 26: **Invalid-answer rates on Video-MMMU.** Qwen3-VL invalid-answer rates (%) are shown without (NR) and with reasoning (R), using 369 shortcut (SC) and 531 non-shortcut questions in every condition. Bold marks the lower rate in each NR/R pair.

Size	Input	SC		Non-SC	
		NR	R	NR	R
2B	Black video	**0.27**	30.08	**0.56**	58.95
	Endpoints	**0.54**	29.00	**0.19**	62.90
	0.02 FPS	**1.63**	18.43	**1.69**	44.63
	0.2 FPS	**1.08**	11.65	**1.13**	31.83
4B	Black video	30.62	**24.93**	52.54	**52.35**
	Endpoints	27.10	**26.02**	**59.13**	65.73
	0.02 FPS	23.31	**22.22**	**49.15**	55.37
	0.2 FPS	21.41	**11.65**	**38.79**	40.49
8B	Black video	**0.81**	17.34	**5.46**	44.07
	Endpoints	**1.63**	11.11	**5.84**	56.50
	0.02 FPS	**2.44**	8.67	**8.47**	37.66
	0.2 FPS	**0.81**	4.88	**6.21**	29.76

Table 27: **Paired shortcut and non-shortcut effects on LVBench.** Qwen3-VL reasoning deltas and paired 95% bootstrap intervals (absolute points) use 193 shortcut (SC) and 1356 non-shortcut questions, with invalid answers counted as wrong. Intervals are unadjusted for multiple comparisons; bold marks the largest delta within each size and question group.

Size	Input	SC Δ	95% CI	Non-SC Δ	95% CI
2B	Black video	+4.15	[−4.15, +12.44]	+0.66	[−1.84, +3.24]
	Endpoints	**+16.58**	[+8.29, +24.87]	−1.77	[−4.28, +0.74]
	0.02 FPS	+12.44	[+5.70, +19.69]	+1.77	[−0.96, +4.50]
	0.2 FPS	+11.92	[+6.22, +18.13]	**+1.92**	[−0.66, +4.50]
4B	Black video	+1.04	[−2.59, +4.66]	+1.55	[+0.37, +2.73]
	Endpoints	**+14.51**	[+8.29, +20.73]	+1.40	[−0.29, +3.17]
	0.02 FPS	+10.36	[+5.18, +15.54]	**+3.54**	[+0.96, +6.19]
	0.2 FPS	+6.74	[+1.55, +11.92]	+2.95	[+0.37, +5.53]
8B	Black video	−20.73	[−27.98, −12.95]	**−7.74**	[−9.66, −5.83]
	Endpoints	−27.98	[−35.23, −20.73]	−7.89	[−10.03, −5.75]
	0.02 FPS	−22.80	[−30.57, −15.03]	−14.75	[−17.33, −12.17]
	0.2 FPS	**−14.51**	[−21.24, −7.77]	−9.66	[−12.32, −6.93]

Table 28: **Invalid-answer rates on LVBench.** Qwen3-VL invalid-answer rates (%) are shown without (NR) and with reasoning (R), using 193 shortcut (SC) and 1356 non-shortcut questions in every condition. Bold marks the lower rate in each NR/R pair.

Size	Input	SC		Non-SC	
		NR	R	NR	R
2B	Black video	**0.00**	1.55	**0.74**	1.92
	Endpoints	**1.04**	5.18	**0.44**	5.68
	0.02 FPS	**0.00**	**0.00**	**0.88**	1.84
	0.2 FPS	**0.00**	0.52	**0.88**	1.03
4B	Black video	2.07	**1.04**	3.91	**0.88**
	Endpoints	10.88	**0.00**	10.25	**2.51**
	0.02 FPS	9.33	**1.55**	12.09	**5.68**
	0.2 FPS	4.15	**1.04**	8.78	**3.83**
8B	Black video	**0.00**	35.75	**0.59**	33.04
	Endpoints	**2.59**	35.23	**3.17**	42.92
	0.02 FPS	**9.84**	40.41	**13.35**	58.92
	0.2 FPS	**0.52**	21.24	**0.29**	29.50

F.3 Denser-Input Comparisons and Aggregate Contrasts

Table 29 reports the available matched 2 FPS comparisons on VideoMME and Video-MMMU. Shortcut gains exceed remaining-question gains in five of six model-by-benchmark conditions at this denser requested input rate. The VideoMME 8B exception remains visible; the available 2 FPS comparison covers these two benchmarks.

Table 30 reports the combined and benchmark-specific question-weighted contrasts with their paired intervals. Table 7 shows how whole-benchmark performance changes with visual input, complementing the group-specific results. Removing detected shortcut questions is an operational stratification, not a certification that the remaining set is shortcut-free. Together these tables expose the heterogeneity underlying the main aggregate finding.

Table 29: **Reasoning with requested 2 FPS inputs, capped at 768 frames.** Qwen3-VL accuracy and invalid-answer rates (%), absolute-point reasoning deltas, and unadjusted paired 95% intervals are reported separately for shortcut (SC) and non-shortcut questions, with bold marking higher accuracy and lower invalid rates within NR/R pairs. No matched LVBench 2 FPS results are available.

Benchmark	Size	Group	n	NR	R	Δ	95% CI	Invalid NR	Invalid R
VideoMME	2B	SC	899	80.87	**88.21**	+7.34	[+4.89, +9.79]	**0.56**	**0.56**
		Non-SC	1801	43.92	**47.08**	+3.16	[+0.89, +5.44]	**0.44**	2.22
	4B	SC	899	89.99	**92.77**	+2.78	[+0.56, +5.01]	**0.11**	1.33
		Non-SC	1801	53.64	**54.36**	+0.72	[−1.44, +2.89]	**0.78**	4.66
	8B	SC	899	**91.10**	85.54	−5.56	[−8.01, −3.11]	**0.11**	8.45
		Non-SC	1801	**53.75**	51.25	−2.50	[−4.66, −0.22]	**1.72**	19.04
Video-MMMU	2B	SC	369	53.12	**74.25**	+21.14	[+15.72, +26.29]	**0.00**	11.92
		Non-SC	531	30.70	**30.89**	+0.19	[−4.33, +4.90]	**0.19**	29.38
	4B	SC	369	68.56	**84.01**	+15.45	[+10.57, +20.05]	23.58	**10.03**
		Non-SC	531	33.15	**39.92**	+6.78	[+2.64, +10.74]	36.72	**30.70**
	8B	SC	369	75.07	**92.68**	+17.62	[+13.28, +21.95]	**0.00**	3.79
		Non-SC	531	38.23	**46.33**	+8.10	[+3.39, +12.81]	**3.20**	23.54

Table 30: **Question-weighted reasoning effects differ by benchmark.** Each question contributes its mean reasoning delta over three Qwen3-VL sizes and four input conditions; SC and non-SC effects and their difference are absolute percentage points. The interaction uses unadjusted question-cluster 95% bootstrap intervals and describes an association, not an internal reasoning mechanism.

Benchmark	n_{SC}	$n_{\mathrm{Non\text{-}SC}}$	SC Δ	Non-SC Δ	Interaction	95% CI
VideoMME	899	1801	+0.80	−3.13	+3.93	[+2.51, +5.36]
Video-MMMU	369	531	+14.23	−3.11	+17.33	[+14.54, +20.13]
LVBench	193	1356	−0.69	−2.34	+1.64	[−0.68, +3.93]
All questions	1461	3688	+3.99	−2.84	+6.83	[+5.68, +7.98]

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The main text reports fixed-group thinking contrasts, paired uncertainty, and joint answer-and-localization results, with scope restricted to the evaluated models and conditions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The conclusion discusses model-dependent shortcut groups and comparison scope. Appendices D and E describe input controls, selection effects and annotation coverage.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The work is empirical and does not present theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Appendices D, E and B specify input controls, scoring, populations, paired resampling and absolute results. Appendix C.5 identifies the evaluation artifacts prepared for release.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: A public release of the evaluation data, code and model outputs with complete access instructions is not yet available. Source videos and model weights remain subject to their original access terms.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Appendix D specifies model controls, frame sampling, generation settings, answer extraction and temporal scoring. No model training is introduced.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Appendix D defines paired question-bootstrap intervals, and Appendix F reports 95% intervals for thinking differences and aggregate contrasts. These intervals quantify question-sampling uncertainty, not variation across inference seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: Appendix C.4 reports token budgets and measured timing profiles, but complete per-experiment hardware, memory and total compute accounting is unavailable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work studies existing video benchmarks and evaluation protocols without deploying a new model. Appendix C.6 discusses inherited data biases, privacy risks and the access terms of source videos.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The main text motivates evidence-dependent evaluation, while Appendix C.6 discusses the benefits of traceable video evidence and risks from overconfidence and surveillance misuse.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The work does not release a new model or high-risk generative system. Its new assets are evaluation questions, annotations and scripts.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No].

    Justification: The evaluated benchmarks and models are cited and source-video access restrictions are noted, but the manuscript does not enumerate all asset-specific licenses and terms. It does not grant redistribution rights for third-party media or model weights.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No].

    Justification: Appendices E, C and C.3 document question generation, behavioral filtering, shortcut categories and human validation. The question data, annotations and accompanying documentation are prepared for release but are not yet publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No].

    Justification: Appendices C.3 and C.1 describe annotation tasks and sample sizes, but complete participant instructions and compensation details are not documented.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No].

    Justification: The manuscript describes human annotation validation but does not document an institutional approval or exemption determination, or a participant risk-disclosure procedure.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes].

Justification: Sections 3 and 4 and Appendix E describe the models used for question generation, behavioral filtering and shortcut classification.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.