KEEP YOUR EYE ON THE CUP: DIAGNOSING IDENTITY AND STATE TRACKING IN VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A wrong answer on a video benchmark can reflect missed motion, confused object identities, or an incorrectly recovered state. We examine these demands with a controlled shell game in which cups swap positions while carrying hidden balls. The benchmark contains 650 synthetic videos and 62 exploratory staged real videos. Its central experiment adds persistent identity markers to the same synthetic videos, holding motion, questions, and answers fixed. On monochrome videos, labeled boxes improve target-cup final-position accuracy by 10.6–65.4 percentage points across eight models. The equal-model gain is 37.5 points for final ball position but only 8.6 points for counting a ball's visits to a position. Accuracy remains lower on longer sequences even with labels. In a separate reconstruction task, only 8.9–30.0% of videos have every reference state correct. Increasing the requested sampling rate from 2 to 4 FPS has mixed effects. These results expose sensitivity to identity information without establishing that every intervening swap was tracked: marked endpoint questions can be answered from initial and final observations.

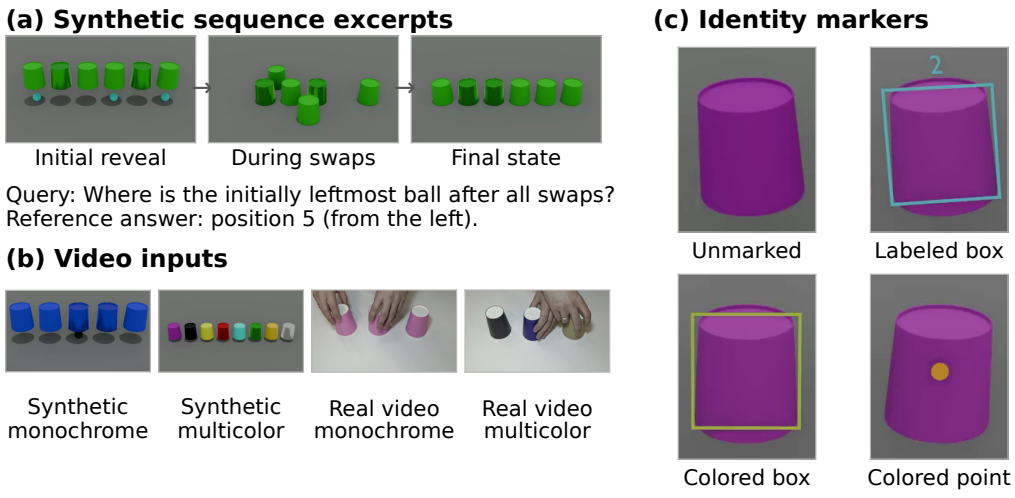


Figure 1: A shell game with controlled identity information. (a) Synthetic sequence excerpts, an endpoint query, and its reference answer. (b) Rendered and recorded inputs; surface-color subsets contain different videos. (c) Marker encodings for three cups in one frame. Labels or persistent colors expose identity; interventions preserve motion and reference answers.

1 INTRODUCTION

Video-language models (VLMs) answer questions about diverse activities and long videos Fu et al. (2025); Wu et al. (2024). Yet an overall score does not reveal whether a model followed the same object through successive movements. We test how answers change when a controlled video supplies persistent identities, reducing the need for motion-based correspondence.

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

In our shell game (Figure 1), cups swap positions while carrying hidden balls; a generator records each state. Recent benchmarks already demonstrate the difficulty of shell-game tracking and intermediate-state reconstruction Liu & Lee (2026); Kwan et al. (2026). We ask a complementary question: *how much does performance improve when the same video supplies persistent cup identities?* We address three research questions: how the benefit depends on the identity cue and question type; how accuracy changes with swap count, including when identities are marked; and whether increasing the requested sampling rate consistently improves answers.

We compare unmarked videos with labeled boxes, colored boxes, and colored points, keeping motion, questions, and reference answers fixed. Endpoint positions, selected intermediate states, and visit counts have different requirements: labels permit matching the initial ball–cup association to the final frame, whereas visit counting refers to the post-swap history. All three research questions use synthetic videos; an exploratory recorded set has different questions and no matched marker intervention.

Labeled boxes help endpoint answers much more than visit counting. Longer sequences remain difficult with labels, and higher requested FPS does not consistently improve accuracy. Our contribution is the matched marker intervention across question requirements, supported by sequence-length and sampling characterization. The model's internal tracking process remains unobserved.

2 Related work

Controlled video diagnostics. Perception Test, MVBench, and TempCompass isolate perceptual and temporal capabilities Pătrăucean et al. (2023); Li et al. (2024); Liu et al. (2024). Programmatic benchmarks such as CLEVRER retain exact latent states Yi et al. (2020). We use this controlled approach to test matched changes to identity information.

Closest shell-game benchmarks. VET-Bench evaluates terminal-position multiple-choice answers on visually indistinguishable objects; its SGCoT method generates coordinate trajectories as intermediate reasoning Liu & Lee (2026). Video-MME-Logical includes hidden-ball cup tricks and shuffling with exact answers and intermediate-state traces Kwan et al. (2026). MET-Bench supplies an initial state and text- or image-encoded swaps, studying length and intermediate reasoning Cohen & Mooney (2026); our inputs are continuous videos. Neither shell games nor reconstruction is new here. We compare persistent identity encodings on the *same* videos and questions, contrasting endpoint and history answers.

State and history diagnostics. VSTAT designs state questions to resist single-frame and visible-end-state shortcuts, and studies uniform frame budgets and temporal stretching Yu et al. (2026). TOC-Bench grounds identity-sensitive, counting, and ordering questions in object tracks and event timelines, filtering text-only, single-frame, and shuffled-frame shortcuts Chen et al. (2026). Our marked endpoints deliberately permit shortcuts to measure the effect of supplied identities; history questions delimit what endpoint success establishes. Sampling sensitivity supports this characterization.

Correspondence and temporal evidence. TAP-Vid and Molmo2 evaluate correspondence or video tracking Doersch et al. (2022); Clark et al. (2026); symbolic entity tracking supplies updates directly Kim & Schuster (2023); Tang et al. (2026). ColorBench motivates caution about color cues Liang et al. (2025). Countix and RepCount test temporal aggregation Dwibedi et al. (2020); Hu et al. (2022), without hidden ball–cup associations.

3 Benchmark and evaluation

3.1 Task and questions

Balls start under cups arranged left to right and stay with their covering cup through pairwise swaps. Given a video and open-ended question, a model returns a scalar, mapping, or sequence without answer choices. Figure 1 shows actual inputs and an example answer.

The six synthetic question schemas ask for a target cup's final position and five aspects of ball state: final position, verification of a selected state, position at a selected step, displacement between two steps, and visit count. The synthetic schemas occupy two categories in the question taxonomy: Tracking (S1) and Temporal reasoning (S2–S6). Perception questions occur only in the recorded split. All ball questions require the initial ball–cup association. Endpoint questions then reference the final state; selected-step and displacement questions reference one or two indexed states; visit counting requires the post-swap history. Each occupied post-swap state counts as a visit, including consecutive occupancies; the initial state is excluded. Markers expose cup identity, but not the swap index or initial association. Appendix Table 6 gives examples and exact schema counts.

In a separate retrospective reconstruction task, the model receives the complete video and returns the cup order at step 0 and after every swap.

3.2 Data and visual conditions

The synthetic set has 650 Blender-rendered videos, 13,628 questions, and durations of 3.75–65 s. Generator metadata records identities, swaps, and reference states. The real set has 62 staged recordings, 434 questions, and durations of 7.7–397.7 s. Each real video has seven questions: five about the initial scene, plus swap count and final cup order. The datasets are unpaired and have different schemas. Independent validation of the real reference answers is not documented. The level designs appear in Appendix A.

Monochrome and multicolored subsets contain different videos and task compositions; their differences are descriptive. The synthetic marker experiment instead adds three overlays to the same videos: labeled cup boxes, boxes with persistent cup-specific colors, and persistent colored points (Figure 1c). Motion, questions, and answers remain fixed. Overlay colors are identity cues, distinct from the unpaired comparison of cup surface colors.

3.3 Protocol and scoring

Open-weight models use vLLM Kwon et al. (2023) and request uniform video sampling at 2 FPS for the main comparisons; the sweep requests 1–4 FPS. These are nominal rates; processed frame counts are incompletely recorded. API models use OpenRouter with provider-native video processing and have no assigned FPS. Videos are 854×480 pixels before model-specific processing. We score one answer for each scheduled video–question pair. Synthetic prompts and questions are in Russian; real-video prompts and questions are in English. Appendix **??** documents the request builder, including a promised initial-order list that it does not append. Appendix C gives English translations where needed.

We report *strict accuracy over all scheduled questions*: a missing, unparseable, or incorrect answer receives zero credit. Mappings and sequences must match the complete reference answer. An entirely unavailable evaluation condition is shown as a dash. For trajectories, we report state accuracy and the fraction of videos with every reference state correct, including step 0; extra predicted step IDs are ignored, so this is an upper bound on exact step-set reconstruction. The mean first-error index conditions on failed trajectories.

Confidence intervals use 10,000 video-cluster bootstrap replicates, keeping matched conditions and models together for each sampled video. They describe variation across videos for the fixed model roster, not repeated inference runs. Model-specific contrasts are exploratory and intervals are unadjusted for multiple comparisons. Appendix A.4 gives scoring and resampling details. Decoding and output limits are not fully documented for every prediction (Appendix A.3), so cross-model differences also reflect configuration and answer-validity differences.

4 Results

4.1 Performance without identity markers

Accuracy depends strongly on what type of information the question requires. For each category of questions model answers can largely vary: for final positions of cup/ball score is within 19.8–60.2% range or 13.5–54.5% respectively (Table 1). Hardest type of question is displacement control

where best model can not achieve 48.5% modal-answer baseline. Overall strict accuracy belongs to 28.2–57.7% interval. These differences motivate schema-specific, answer-prior-calibrated analysis. Recorded-video results appear in Appendix Table 8.

Table 1: Synthetic accuracy by question schema. Strict accuracy (%) on 13,628 questions from 650 unmarked videos. Open-weight models request 2 FPS; API models use provider-native sampling.

Model	Tracking	Temporal reasoning					All
	Cup final S1	Verify S2	Visits S3	Ball at step S4	Displacement S5	Ball final S6	
Qwen3.5-4B	19.8	60.1	46.6	10.8	28.3	13.5	28.2
Qwen3.5-9B	42.8	60.6	49.6	29.5	39.7	32.8	42.6
Qwen3.5-27B	59.4	72.0	51.9	54.8	46.0	53.3	56.8
Qwen3.5-35B-A3B	55.7	68.8	49.2	50.8	40.7	47.2	52.7
Qwen3.5-122B-A10B	55.1	62.1	49.0	48.8	39.7	48.4	51.3
Qwen3.5-397B-A17B	57.8	66.6	52.3	54.6	41.9	54.5	55.2
Gemma-4-31B-it	57.0	68.6	48.0	53.7	45.6	46.0	53.8
InternVL3.5-38B	35.4	51.8	29.8	29.5	23.5	22.6	32.7
Seed-2.0-Lite	58.2	73.8	55.0	59.5	43.9	54.0	57.7
Gemini-3-Flash	54.0	55.3	41.7	31.1	30.3	26.0	42.1
Gemini-3.5-Flash	60.2	56.8	49.5	47.2	46.1	42.2	52.0
Modal answer (LOVO)	15.9	55.4	53.4	16.2	48.5	16.6	31.3
Uniform (metadata)	15.9	50.0	16.7	16.4	9.1	16.4	20.1

Missing or unparseable answers count as wrong. Baselines use leave-one-video-out answer frequencies or scene metadata (Appendix B.1); Table 6 details the question types.

4.2 Identity cues help endpoints more than history

Persistent labels help endpoint retrieval much more than visit history. On matched monochrome videos at 2 FPS, the equal-model gains are 37.5 points (95% interval: 35.5–39.6) for final ball position and 8.6 points (7.5–9.8) for visit counting (Figure 2); their direct difference is 28.9 points (26.8–31.0).

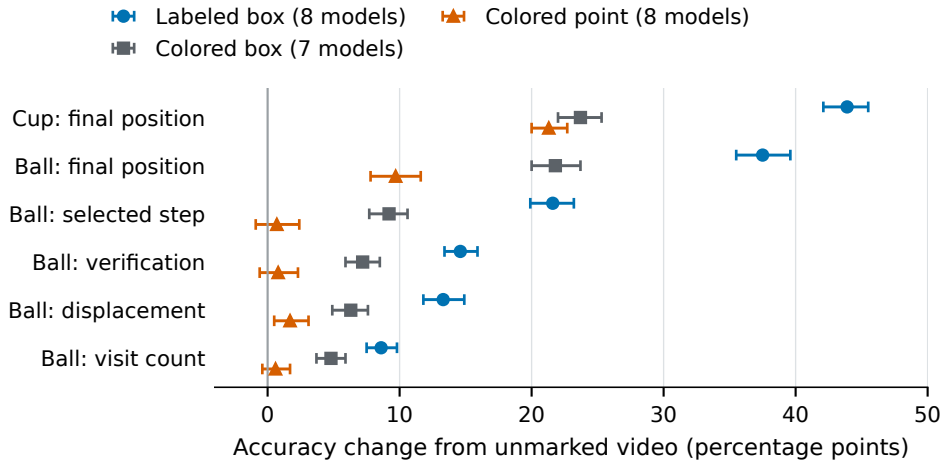


Figure 2: Identity cues help endpoints more than history. Paired marker-minus-unmarked strict accuracy on monochrome videos at 2 FPS. Points are equal-model means; bars are 95% video-cluster intervals. Labeled boxes and points use eight models; colored boxes use seven (397B unavailable).

The visit gain does not by itself demonstrate use of the video. Final-position questions can exploit persistent identities without recovering every swap, whereas visits require occupancy across post-swap states. Yet their answer priors also differ: a leave-one-video-out modal baseline scores 17.5% on monochrome final-ball endpoints and 50.4% on visits, exceeding the labeled models' 49.5% visit mean (Appendix B.1).

Labeled boxes help all eight models in the matched comparison. Every labeled-box contrast in Table 2 is positive in both question groups, with intervals excluding zero. Target-cup gains span 10.6–65.4 points and pooled ball-state gains span 2.3–31.6 points. The pooled group contains endpoints and history questions, so it is broader than visit counting alone.

Table 2: Per-model effects of supplying cup identities. Matched monochrome videos at 2 FPS: 2,480 target-cup and 6,608 ball-state questions per model over 400 videos. Unmarked values are strict accuracy (%); marker columns are paired changes (pp), with 95% video-cluster intervals. Missing and invalid answers remain wrong; the 397B colored-box condition is unavailable.

Target-cup final position

Model	Unmarked	Labeled box Δ	Colored box Δ	Colored point Δ
Qwen3.5-4B	10.5 [9.1, 12.0]	+24.8 [+22.4, +27.2]	+10.9 [+8.8, +13.0]	+6.9 [+4.9, +8.9]
Qwen3.5-9B	23.6 [21.6, 25.7]	+40.5 [+37.7, +43.4]	+18.7 [+16.1, +21.4]	+13.5 [+10.8, +16.1]
Qwen3.5-27B	35.4 [32.8, 38.1]	+56.0 [+52.9, +59.1]	+40.7 [+37.6, +43.8]	+40.8 [+37.9, +43.6]
Qwen3.5-35B-A3B	31.8 [29.3, 34.3]	+55.2 [+52.5, +57.9]	+32.9 [+29.8, +35.9]	+39.6 [+36.8, +42.6]
Qwen3.5-122B-A10B	28.9 [26.2, 31.5]	+57.6 [+54.9, +60.3]	+46.0 [+43.1, +49.0]	+46.9 [+44.1, +49.6]
Qwen3.5-397B-A17B	31.2 [28.5, 34.1]	+65.4 [+62.8, +68.0]	—	+4.8 [+2.7, +6.9]
Gemma-4-31B-it	33.4 [30.8, 36.2]	+40.6 [+37.2, +44.2]	+16.2 [+13.4, +19.1]	+16.8 [+13.9, +19.7]
InternVL3.5-38B	27.3 [25.0, 29.7]	+10.6 [+8.3, +12.9]	+0.1 [-1.7, +1.8]	+1.3 [-0.5, +3.1]

Ball-state questions

Model	Unmarked	Labeled box Δ	Colored box Δ	Colored point Δ
Qwen3.5-4B	27.4 [26.1, 28.7]	+12.2 [+10.6, +13.7]	+5.1 [+3.8, +6.5]	-0.3 [-1.6, +1.0]
Qwen3.5-9B	35.2 [33.6, 36.8]	+11.7 [+10.1, +13.3]	+6.1 [+4.7, +7.6]	-2.6 [-3.9, -1.3]
Qwen3.5-27B	45.3 [43.0, 47.6]	+25.0 [+23.0, +27.0]	+17.0 [+15.2, +18.9]	+3.8 [+1.4, +6.1]
Qwen3.5-35B-A3B	40.8 [38.7, 43.0]	+23.7 [+21.6, +25.8]	+11.6 [+9.8, +13.4]	+3.1 [+1.0, +5.3]
Qwen3.5-122B-A10B	36.8 [33.9, 39.8]	+30.5 [+28.7, +32.4]	+24.8 [+23.0, +26.7]	+16.1 [+13.9, +18.4]
Qwen3.5-397B-A17B	41.9 [38.8, 45.1]	+31.6 [+29.7, +33.5]	—	+0.8 [-0.4, +2.0]
Gemma-4-31B-it	43.1 [40.9, 45.5]	+16.7 [+14.4, +19.1]	+5.6 [+3.8, +7.4]	+2.9 [+0.9, +4.9]
InternVL3.5-38B	31.6 [29.9, 33.4]	+2.3 [+1.0, +3.6]	-1.0 [-2.2, +0.3]	-1.9 [-3.3, -0.5]

Cue design can reverse the direction of the effect. Labeled boxes have the largest reported point gain for every model and question group. Colored-point target-cup estimates are positive throughout, but pooled ball-state effects include clear gains, estimates whose intervals include zero, and losses. The encodings change symbol content, size, and salience together with identity information, and labeled-box gains do not consistently extend to matched multicolored videos (Appendix Table 10).

The larger endpoint gain persists among valid answers. All contrasts retain missing or unparseable answers as wrong. Restricting to answers valid in both conditions still gives a 29.0-point difference between endpoint and visit gains (26.9–31.1). This is a selection-conditioned check, not a corrected ability estimate (Appendix B.2).

4.3 LONGER SEQUENCES REMAIN DIFFICULT WITH IDENTITY LABELS

Identity labels do not remove the association between length and error. For every model, ball-state accuracy is lower in the 30–50-swap group than in the 1–3-swap group, both without markers and with labeled boxes (Figure 3).

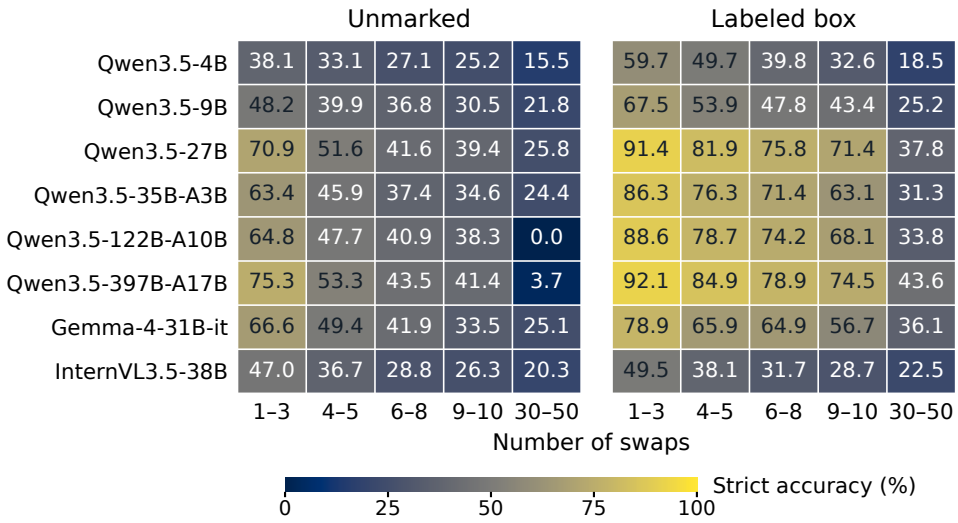


Figure 3: Labels help, but accuracy is lower on longer sequences. Strict accuracy (%) on matched monochrome ball-state questions at 2 FPS, with a common color scale across conditions. The 30–50-swap group is a separate design level; no videos have 11–29 swaps. Groups contain different videos and question mixtures. The low long-sequence baselines for Qwen-122B/397B include missing answers, which remain incorrect. Appendix Figure 6 shows intervals; Table 11 gives sample counts.

Longer sequences remain harder with cup count and question type fixed. This comparison holds cup count, question schema, and answer space fixed and adjusts for model with fixed effects. Its 2,560 matched model–question pairs from 70 videos and eight models cover 1–47 swaps. The linear-probability model estimates a decline of 0.8 points per additional swap without markers (95% video-cluster interval: 0.6–1.1) and 1.3 points with labels (1.1–1.5). Excluding the two models with incomplete baseline records retains declines of 0.7 and 1.2 points per swap. Labeled accuracy remains higher in every observed bin, while the fitted advantage narrows with swap count. Duration and input length remain uncontrolled, so this establishes an association under these conditions rather than an isolated state-maintenance failure or out-of-distribution length result.

Most full trajectory reconstructions fail. Only 8.9–30.0% of the 650 videos have every reference state reconstructed correctly (Table 3). Among failed trajectories, the mean first-error step is 1.5–2.4, and errors can occur at step 0. Because this mean excludes successful videos, it measures where failed outputs first disagree rather than an unconditional tracking horizon.

Table 3: Most reference-state sequences contain errors. Trajectory scoring at 2 FPS on 650 videos. All states correct requires every reference state, including step 0, to match; extra predicted step IDs are ignored. First error averages the first incorrect state index among failed trajectories only. The 400 monochrome and 250 multicolored videos are unpaired and differ in length.

Model	All states correct (%)	First error	Monochrome all states (%)	Multicolored all states (%)
Qwen3.5-4B	11.7	1.5	4.2	23.6
Qwen3.5-9B	10.6	1.7	4.2	20.8
Qwen3.5-27B	24.6	2.4	8.2	50.8
Qwen3.5-35B-A3B	20.8	2.2	8.0	41.2
Qwen3.5-122B-A10B	25.7	2.2	7.5	54.8
Qwen3.5-397B-A17B	30.0	2.3	11.8	59.2
Gemma-4-31B-it	26.0	2.0	10.5	50.8
InternVL3.5-38B	8.9	1.7	7.8	10.8

Higher multicolored scores do not isolate a benefit of color. Every model reconstructs at most 11.8% of monochrome sequences correctly, while multicolored accuracy is higher for all eight models. The subsets contain different videos, however, and multicolored sequences end by swap 10. Figure 4 therefore separates accuracy at one eligible step from full-prefix correctness without treating the color gap as a controlled effect. Per-model curves and cup-count analyses appear in Appendix B.5.

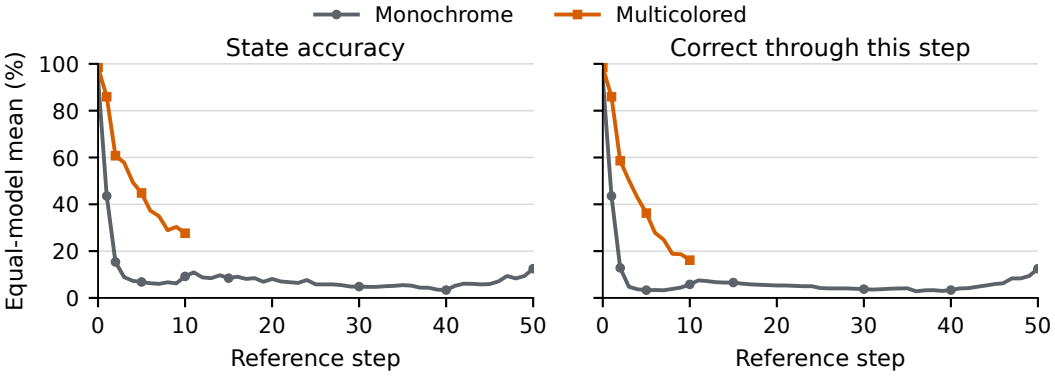


Figure 4: Most reference-state sequences contain errors. Retrospective reconstruction at 2 FPS, averaged equally over eight models. State accuracy scores one cup order; prefix correctness requires every state through that step. Later steps contain fewer, different videos, as the counts show. These are not measurements of online updates or within-video error accumulation. Multicolored trajectories end by step 10.

4.4 Higher requested sampling rates have mixed effects

Moving from 1 to 2 FPS usually improves accuracy, but only modestly. From 1 to 2 FPS, paired changes range from -0.5 to $+4.9$ points, with five of six estimates positive (Table 4; Figure 5). The intervention can change both temporal evidence and input length.

Table 4: Higher requested sampling rates have model-dependent effects. Nominal uniform sampling, with 13,628 scheduled questions across 650 synthetic videos per available cell. FPS columns are strict accuracy (%); changes are paired pp with 95% video-cluster intervals for six models evaluated at every rate.

Model	1 fps	2 fps	3 fps	4 fps	Δ(2–1)	Δ(4–2)
Qwen3.5-4B	28.7	28.2	26.1	25.8	-0.5 [-1.5, +0.4]	-2.4 [-3.4, -1.5]
Qwen3.5-9B	40.3	42.6	42.3	44.0	+2.3 [+1.3, +3.2]	+1.4 [+0.5, +2.4]
Qwen3.5-27B	52.0	56.8	56.2	57.5	+4.9 [+3.9, +5.8]	+0.7 [-0.2, +1.5]
Qwen3.5-35B-A3B	47.9	52.7	52.0	53.2	+4.9 [+3.8, +5.9]	+0.5 [-0.4, +1.4]
Qwen3.5-122B-A10B	50.7	51.3	56.4	56.3	+0.6 [-0.7, +2.0]	+5.0 [+3.9, +6.1]
Gemma-4-31B-it	52.6	53.8	51.8	51.1	+1.3 [+0.2, +2.3]	-2.7 [-3.7, -1.7]

Higher rates have heterogeneous effects and no shared best rate. For the 2-to-4 FPS contrast, two models have positive intervals, two have negative intervals, and two intervals include zero. No single rate maximizes all six complete sweeps, so a sampling choice that benefits one model can reduce another's accuracy.

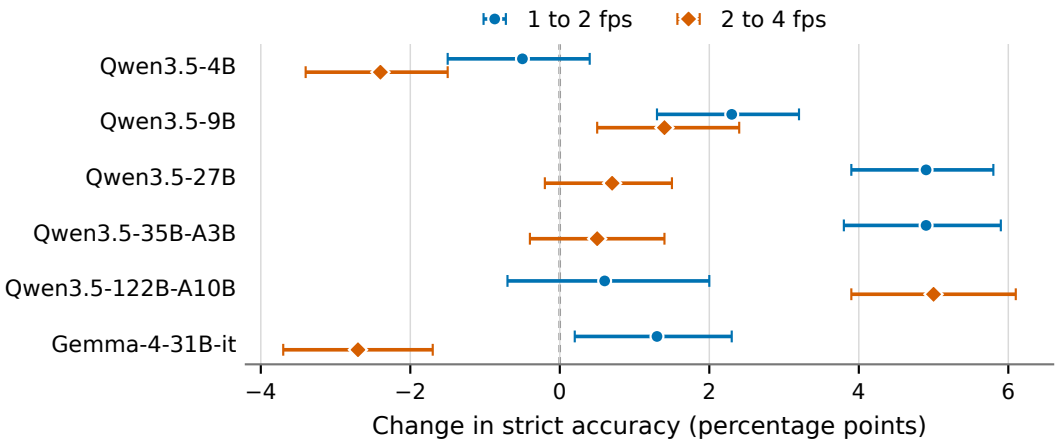


Figure 5: Higher requested FPS does not consistently improve accuracy. Paired changes in strict accuracy for six models evaluated at all four rates, each with 13,628 scheduled questions over 650 videos. Bars are 95% video-cluster intervals; missing and invalid answers remain wrong. Table 4 gives absolute scores and available conditions for two additional models.

Sampling gains can include improved answer validity. Qwen3.5-122B-A10B rises 5.0 points (3.9–6.1) from 2 to 4 FPS while validity rises from 85.0% to 100.0%; among answers valid at both rates, the gain is 1.4 points. Treating absent records as unknown gives bounds of $[-10.0, +5.0]$ (Appendix B.4).

Long videos remain difficult at every available sampling rate. Every model is below 40.0% accuracy in the longest duration group at each available rate (Appendix Figure 7), indicating the remaining challenges of long-horizon state tracking tasks.

4.5 Model size and cross-benchmark correlations

Parameter correlations are positive but uncertain. Within Qwen3.5, both total and active parameter counts correlate positively with overall strict accuracy (Table 5), but all four tests have $p > 0.20$. The six-model synthetic and five-model recorded comparisons mix dense and MoE architectures; 397B has no recorded result.

Table 5: Pearson correlations with shell-game accuracy; n is the model count. Parameter rows use overall strict accuracy at 2 FPS and nominal raw billions within Qwen3.5. Benchmark rows use the best numeric-question strict score and best parsed-conditional external score.

Comparator	Synthetic r (n)	Recorded r (n)
Nominal total parameters (B)	+0.445 (6)	+0.616 (5)
Nominal active parameters (B)	+0.602 (6)	+0.605 (5)
Countix exact	+0.737 (10)	+0.427 (10)
Countix off-by-one	+0.808 (10)	+0.491 (10)
RepCount exact	+0.899 (10)	+0.772 (10)
RepCount off-by-one	+0.811 (10)	+0.705 (10)

Observed correlations are higher on synthetic videos and with RepCount. We compare the best available numeric-question scores with Countix (Dwibedi et al., 2020) and RepCount (Hu et al., 2022), selecting settings separately for each dataset and metric. Numeric schemas (S1, S3–S6; R1–R3) include positions and signed displacements as well as counts. External accuracy conditions on parsed answers; off-by-one allows $|\hat{y} - y| \leq 1$. Counting unparsed answers as wrong and requiring complete evaluated grids retains higher synthetic correlations ($r = 0.827$–0.923) than recorded correlations (0.421–0.784), with 10–11 models. Selected settings, response coverage, and small, family-clustered rosters limit these associations: they establish neither a scaling law nor benchmark transfer.

5 Discussion and conclusion

Identity cues help endpoint retrieval more than history recovery. All eight models benefit from labeled boxes, yet endpoint questions improve much more than visit counting, and colored-point effects on pooled ball-state questions include gains, uncertain effects, and losses. Recognizable identities therefore help without guaranteeing correct sequence reasoning.

Long sequences expose failure beyond endpoint accuracy. Accuracy is lower on longer sequences even with labels, and the fixed nine-cup analysis preserves this association. Only 8.9–30.0% of videos have every reference state correct. These findings separate endpoint success from reliable reconstruction without identifying the model's internal failure mechanism.

More requested frames are not a reliable remedy. Higher requested FPS produces gains for some models and losses for others, so sampling rate should be reported and evaluated rather than treated as an automatic improvement. The exploratory recorded videos add an input setting with two dynamic schemas; they do not provide a matched synthetic–real or marker comparison.

Scope and limitations. Markers change identity, localization, symbol content, and salience together. The observed gains therefore cannot be assigned solely to better temporal tracking. Sequence length covaries with duration and input length, and trajectory reconstruction requires longer outputs for longer sequences. The synthetic scenes are restricted; the real set is small and lacks documented independent answer validation. Human performance and a vision-based tracking baseline have not been measured. Strict accuracy includes inference and formatting failures, decoding provenance is incomplete, and uncertainty describes sampled videos rather than repeated model runs.

The benchmark supports intervention-based diagnosis. Progress can be measured as more reliable answers under controlled changes to identity information. The evidence separates endpoint success from history-dependent answers and reference-state reconstruction while keeping the mechanism of improvement open to further tests.

Reproducibility statement

We plan to release the videos, generator, annotations, evaluation code, and available run configurations. Section 3.3 and Appendix A.3 describe the limits on inference reproducibility.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing boilerplate Python code and auxiliary scripts for experiments, formatting and stylistic editing of the manuscript, and reviewing the literature for the Related Work section. We did not use generative AI tools to formulate research problems, design experiments, collect or annotate data, or formulate the main scientific conclusions. We reviewed all AI-assisted work and take full responsibility for the final content of this paper and its accompanying code, including material produced with the assistance of generative AI.

REFERENCES

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026. URL https://arxiv.org/abs/2607.00248.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A temporal object consistency benchmark for video large language models, 2026. URL https://arxiv.org/abs/2605.09904v2.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28652–28668, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Clark_Molmo2_Open_Weights_and_Data_for_Vision-Language_Models_with_Video_Understanding_CVPR_2026_paper.html.

Vanya Cohen and Raymond Mooney. MET-Bench: Multimodal entity tracking for evaluating the limitations of vision-language and reasoning models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, Seoul, South Korea, July 2026. PMLR. URL https://icml.cc/virtual/2026/poster/62633.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13610–13626. Curran Associates, Inc., 2022. doi: 10.52202/068431-0989. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10387–10396, 2020. doi: 10.1109/CVPR42600.2020.01040. URL https://doi.org/10.1109/CVPR42600.2020.01040.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhui Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Google DeepMind. Gemini 3 flash model card. Google DeepMind model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Google DeepMind. Gemma 4 model card. Google AI for Developers, 2026. URL https://ai.google.dev/gemma/docs/core/model_card_4.

Huazhang Hu, Sixun Dong, Yiqun Zhao, Dongze Lian, Zhengxin Li, and Shenghua Gao. TransRAC: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19013–19022, 2022. doi: 10.1109/CVPR52688.2022.01843. URL https://doi.org/10.1109/CVPR52688.2022.01843.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3835–3855, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.213. URL https://aclanthology.org/2023.acl-long.213/.

Hohin Kwan, Hongyu Li, Ray Zhang, Manyuan Zhang, Xianghao Kong, Anyi Rao, Jiahao Xie, and Si Liu. Video-MME-Logical: A controlled diagnostic benchmark for video temporal-logical reasoning, 2026. URL https://arxiv.org/abs/2606.27828v1. Accepted to Findings of EMNLP 2026.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Yijun Liang, Ming Li, Chenrui Fan, Ziyue Li, Dang Nguyen, Kwesi Cobbina, Shweta Bhardwaj, Jiuhai Chen, Fuxiao Liu, and Tianyi Zhou. ColorBench: Can VLMs see and understand the colorful world? a comprehensive benchmark for color perception, reasoning, and robustness. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4803. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d34497330b1fd6530f7afd86d0df9f76-Abstract-Datasets_and_Benchmarks_Track.html.

Tiedong Liu and Wee Sun Lee. Can vision-language models solve the shell game?, 2026. URL https://arxiv.org/abs/2603.08436.

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? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.517. URL https://aclanthology.org/2024.findings-acl.517/.

OpenGVLab. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

OpenRouter. Openrouter API documentation. OpenRouter developer documentation, 2026. URL https://openrouter.ai/docs/quickstart.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761. Curran Associates, Inc., 2023. doi: 10.52202/075280-1852. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8540fba4abdc7f9f7a7b1cc6cd60e409-Abstract.html.

Qwen Team. Qwen3.5-122B-A10B model card and checkpoint configuration. Hugging Face model repository, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-122B-A10B/blob/main/config.json.

Qwen Team. Qwen3.5-27B model card and checkpoint configuration. Hugging Face model repository, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-27B/blob/main/config.json.

Qwen Team. Qwen3.5-397B-A17B model card and checkpoint configuration. Hugging Face model repository, 2026c. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B/blob/main/config.json.

Qwen Team. Qwen3.5-4B model card and checkpoint configuration. Hugging Face model repository, 2026d. URL https://huggingface.co/Qwen/Qwen3.5-4B/blob/main/config.json.

Qwen Team. Qwen3.5-9B model card and checkpoint configuration. Hugging Face model repository, 2026e. URL https://huggingface.co/Qwen/Qwen3.5-9B/blob/main/config.json.

Qwen Team. Qwen3.5-35B-A3B model card and checkpoint configuration. Hugging Face model repository, 2026f. URL https://huggingface.co/Qwen/Qwen3.5-35B-A3B/blob/main/config.json.

Zilu Tang, Qiao Zhao, Gabriel Franco, Derry Wijaya, Aaron Mueller, Sebastian Schuster, and Najoung Kim. Do language models track entities across state changes? In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, Seoul, South Korea, July 2026. PMLR. URL https://icml.cc/virtual/2026/poster/64207.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxYzANYDB.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, and Saining Xie. Benchmarking visual state tracking in multimodal video understanding, 2026. URL https://arxiv.org/abs/2606.03920v1.

A Data and evaluation details

A.1 Question types

Table 6 defines every question schema, its state tracking requirements, answer form, and amount in the benchmark. The inference prompts appear in Appendix C.

Table 6: Complete question inventory. State dependency names the latent scene information referenced by a question; it does not imply that the information is visible in one frame. In the final column, N is the number of scheduled questions in one complete run.

Code	Example question	State dependency	Answer / N
Synthetic: 650 videos, 13,628 questions			
S1	Where does the third white cup finish?	Initial cup identity; final cup state	Integer / 4,100
S2	After swap 7, is the yellow ball at position 8?	Initial ball–cup association; indexed cup state	Yes/no / 1,930
S3	How many times does the yellow ball visit position 6?	Initial association; full state sequence[†]	Integer / 1,930
S4	Where is the yellow ball after swap 2?	Initial association; indexed cup state	Integer / 1,930
S5	What is the ball's signed displacement from swap 5 to 6?	Initial association; two indexed cup states	Signed integer / 1,808
S6	Where does the yellow ball finish?	Initial association; final cup state	Integer / 1,930
Real: 62 videos, 434 questions			
R1	How many cups are there?	Initial scene	Integer / 62
R2	How many balls are there?	Initial scene	Integer / 62
R3	How many swaps occur?	Complete event sequence	Integer / 62
R4	Which cup starts above each ball?	Initial ball–cup association	Mapping / 62
R5	What color is each ball?	Initial scene	Mapping / 62
R6	What color is each cup?	Initial scene	Mapping / 62
R7	What is the final left-to-right cup order?	Initial cup identities; final cup state	Sequence / 62

[†]A visit is counted for every post-swap state in which the ball occupies the target position. The initial state is excluded; repeated occupancy after successive swaps is counted repeatedly.

A.2 Dataset design

Table 7 combines the synthetic and recorded level designs. The synthetic corpus has 400 monochrome and 250 multicolored videos. Its metadata supplies identities, colors, swap sequences, and reference states. Only the monochrome subset includes the 30–50-swap level; the multicolored subset ends at 10 swaps. Color comparisons therefore also compare different difficulty distributions. All marker and FPS comparisons use the same 650 unique evaluation videos; a separate development split is not specified.

The real set has 39 monochrome and 23 multicolored videos. Only swap count (R3) and final cup order (R7) query dynamic quantities; the other five questions concern the initial scene. Its overall score is not an accuracy measure on the synthetic ball-state tasks. Independent validation of real reference answers is not documented.

Table 7: Dataset design by level. Surface colors, object counts, swaps, and length vary jointly; color subsets are not paired interventions. Synthetic levels have 10 videos per level×cups×balls cell.

Level	Videos	Cups	Balls	Swaps	Cup colors	Duration (s) Median	Max
Synthetic: 650 videos, 13,628 questions							
L1	80	2–9	1	1–10	one	8.8	15
L2	170	3–9	2–6	1–10	one	8.8	15
L3	70	3–9	3–9	1–10	one	9.4	15
L4	80	2–9	1	1–10	multiple	8.8	15
L5	170	3–9	2–5	1–10	multiple	10	15
L6	80	2–9	1–7	30–50	one	52.5	65
Recorded: 62 videos, 434 questions							
L0	3	2	1	1	one	10.4	11.7
L1	15	3	1	1–3	one	14.7	20.3
L2	17	3–4	1	3–5	multiple	20.8	29
L3	14	3–5	2	5–10	one	31.8	47.5
L4	6	4–6	2–3	12–20	multiple	55.9	82.4
L5	3	6	3	37–39	one	144.2	156.1
L6	4	5–8	3–5	39–163	one	173.3	397.7

A.3 Models and decoding settings

Table 1 lists the primary synthetic QA results. Model and provider documentation describes the architectures and interfaces OpenGVLab (2025); Qwen Team (2026d;e;b;f;a;c); Google DeepMind (2026; 2025); ByteDance Seed (2026); OpenRouter (2026). Open-model processing configurations specify TorchCodec decoding and uniform sampling at the requested rate.

The one-question configurations specify seed 0 and model-specific context and decoding limits. Representative Qwen and InternVL configurations specify temperature 0.6; Gemma configurations specify 1.0. These files commonly give a 20,000-token output cap and 14,000-token thinking budget, whereas the experiment notes give a 2,000-token output cap. Predictions are not linked to configuration versions, so the applicable limit cannot be established for every result. Processed-frame counts, provider-side extraction, and output truncation are also incompletely documented. InternVL configurations name an R1-style system prefix, whose use cannot be verified for every prediction. The retained QA exports permit re-computing scores; trajectory exports preserve derived correctness rows rather than raw answers. Neither reconstructs every inference configuration.

A.4 Scoring and uncertainty

Question answering. For N scheduled questions, strict accuracy is

$$\mathrm{Acc}_{\mathrm{strict}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i]. \tag{1}$$

Every missing or invalid answer contributes zero. Scalars are compared after normalizing case, numeric notation, yes/no variants, and object-ID underscores. Mappings require exact equality of the normalized key–value dictionaries; sequences require the same length and order. The validity rate is the fraction of scheduled questions with a nonempty parsed value for the requested question; it measures parser coverage, not correctness or complete type compliance. Record availability is separate: an existing record may contain an inference error or an unparseable response. Neither missingness nor invalidity alone identifies its cause. An unavailable evaluation condition is a dash, not a zero. Matching scheduled questions determines eligibility for paired comparisons, regardless of response validity. Accuracies are percentages and differences are percentage points.

Trajectories. The scorer creates a row for every reference state, including step 0, and indexes predictions by their declared step number. The complete normalized cup sequence must match in length and order. Missing states, malformed trajectories, and wrong sequences are incorrect. For duplicate step numbers, the last occurrence is used; predicted IDs absent from the reference are ignored.

Thus, a video may have every reference state correct despite extra predicted step IDs. Raw predicted step sets are absent from the retained export, so this leniency cannot be quantified retrospectively. The reported success rate is an upper bound on success requiring an exact set of reference step IDs. The first-error mean includes only failed trajectories and may equal zero. It is not an unconditional time-to-failure estimate.

Resampling. Confidence intervals are percentile video-cluster intervals from 10,000 replicates. Videos are sampled with replacement and all questions from each sampled video remain together. Matched conditions are resampled jointly; model averages keep every model's responses to a sampled video together. The fixed-stratum regression is refitted in each replicate. Marker and swap-bin analyses use seed 20270914; regression uses 20270915. Other analysis-specific seeds are recorded in the accompanying code. Intervals measure variation across sampled videos, not repeated stochastic runs.

B Additional results

This appendix gives per-schema estimates, sequence-length details, color controls, and additional diagnostics behind the main comparisons. Unless indicated otherwise, accuracy and validity are percentages and changes are percentage points. Strict accuracy always uses the scheduled-question denominator: a missing record and an available record with an invalid or unparseable answer are distinct outcomes, but both receive zero credit. A dash denotes an unevaluated condition rather than a zero score.

B.1 Unmarked question answering

Main Table 1 gives detailed synthetic results; Table 8 reports the exploratory recorded-video results. Schema codes are defined in Table 6. The datasets have different questions and denominators; only R3 and R7 in the recorded set query changes during the video. Gemini-3.5 scores 71.0% overall on the recorded set, close to the 71.9% maximum, with 92.9% answer validity.

Table 8: Exploratory recorded-video strict accuracy (%). Each run has 434 scheduled questions across 62 videos, with 62 questions per schema. R3 and R7 query video dynamics; the other five schemas concern the initial scene.

Model	R1	R2	R3	R4	R5	R6	R7	All	Valid	No record
Qwen3.5-4B	98.4	77.4	12.9	37.1	41.9	79.0	25.8	53.2	99.1	0.0
Qwen3.5-9B	96.8	79.0	33.9	35.5	61.3	87.1	37.1	61.5	99.5	0.0
Qwen3.5-27B	98.4	71.0	66.1	56.5	69.4	85.5	56.5	71.9	99.5	0.0
Qwen3.5-35B-A3B	98.4	80.6	40.3	41.9	66.1	87.1	53.2	66.8	99.1	0.0
Qwen3.5-122B-A10B	100.0	87.1	50.0	51.6	69.4	88.7	48.4	70.7	100.0	0.0
Gemma-4-31B-it	69.4	66.1	51.6	56.5	66.1	66.1	50.0	60.8	70.5	0.0
InternVL3.5-38B	9.7	4.8	8.1	0.0	0.0	0.0	0.0	3.2	41.0	0.0
Seed-2.0-Lite	100.0	62.9	54.8	54.8	59.7	79.0	56.5	66.8	99.5	0.0
Gemini-3-Flash	96.8	64.5	54.8	43.5	59.7	74.2	45.2	62.7	95.9	0.0
Gemini-3.5-Flash	95.2	67.7	69.4	59.7	61.3	79.0	64.5	71.0	92.9	0.0

All averages the seven schemas. Valid is the share of scheduled questions with a parseable, nonempty answer; No record is the share without a prediction record. Missing and invalid answers count as wrong. Different videos and schemas prevent direct comparison with synthetic scores.

Answer-prior calibration. For each schema, the leave-one-video-out (LOVO) baseline predicts the most frequent reference answer in all other videos, breaking ties toward the smaller answer. It uses neither video nor question text. Each of the 13,628 references is counted once, independent of model and sampling rate. The baseline reaches 31.3% overall; verification (S2), visits (S3), and displacement (S5) reach 55.4%, 53.4%, and 48.5%. Their full-corpus modal answers are true, zero, and zero. For the 400 monochrome videos used in the marker analysis, LOVO scores are 17.5% on final ball position and 50.4% on visits. These are within-corpus calibrations, not generalization results on an independent generator. The accompanying analysis exports every answer frequency. Different priors and answer spaces limit interpretations of cross-schema score gaps.

Scene-aware reference guesser. The synthetic table gives the analytic expected accuracy of a uniform guesser that knows cup count C and swap count S: positions are uniform over $1, \ldots, C$, displacements over $-(C-1), \ldots, C-1$, yes/no over two values, and visits over $0, \ldots, S$. It reaches 20.1% overall and 16.7% on visits. This privileged metadata baseline is not a vision model. The earlier Monte Carlo guesser capped visits at 10 and scored 18.9% overall; the new analytic baseline covers the valid answers above 10 as well (80 of 1,930 visit questions; maximum 25).

For real questions, the existing metadata-aware guesser scores 24.2% strict and 29.8% partial on 434 questions (200 draws, seed 0). It knows cup count and mapping keys, samples ball counts from 1 to C, mapping values independently, final cup order uniformly, colors from the ten-color prompt palette, and swap counts from 1–39. Partial scoring gives scalar answers zero or one credit and averages correct keys or positions for mappings and sequences before averaging over questions.

B.2 Identity-marker interventions

Table 2 gives per-model baselines and paired changes for all three marker encodings on monochrome synthetic videos. Every available cell has 9,088 scheduled questions over the same 400 videos. Most cells have 99.8–100.0% record availability and 98.8–100.0% answer validity. The material exceptions are the 122B unmarked condition (77.6%/77.6% availability/validity), the 397B unmarked condition (89.4%/80.5%), and the 397B colored-point condition (77.6%/77.6%); the 397B colored-box condition was not evaluated. Thus, paired strict-accuracy changes can include changes in availability or parseability as well as answer correctness.

Table 9 provides the distinct schema-level estimates and video-cluster intervals. It separates endpoint questions from questions about a selected state, displacement, or visit history; a correct endpoint does not establish that every intervening swap was tracked.

Table 9: Marker effects by synthetic question schema. The unmarked, labeled-box, and colored-point estimates give equal weight to eight models; colored-box changes use seven because no 397B colored-box results are available. Each bootstrap resamples the same videos with all eligible models and conditions retained. Values are strict accuracy (%) or paired pp with 95% intervals.

Question schema	Unmarked	Labeled box Δ	Colored box Δ	Colored point Δ	Videos
Target-cup final position	27.8 [25.9, 29.8]	+43.9 [+42.1, +45.5]	+23.7 [+22.0, +25.3]	+21.3 [+20.0, +22.7]	400
Ball final position	26.4 [24.3, 28.5]	+37.5 [+35.5, +39.6]	+21.8 [+20.0, +23.7]	+9.7 [+7.8, +11.6]	400
Ball position at a step	32.5 [30.3, 34.8]	+21.6 [+19.9, +23.2]	+9.2 [+7.7, +10.6]	+0.7 [-0.9, +2.4]	400
Ball visits to a position	40.9 [37.9, 44.0]	+8.6 [+7.5, +9.8]	+4.8 [+3.7, +5.9]	+0.6 [-0.4, +1.7]	400
Ball-position verification	57.4 [55.3, 59.5]	+14.6 [+13.4, +15.9]	+7.2 [+5.9, +8.5]	+0.8 [-0.6, +2.3]	400
Ball displacement	31.0 [29.0, 33.1]	+13.3 [+11.8, +14.9]	+6.3 [+4.9, +7.6]	+1.7 [+0.5, +3.1]	370

Direct contrast and validity sensitivity. Using the same eight models and 400 monochrome videos, a joint video bootstrap gives a 28.9-point difference between the labeled-box gain on final ball position and on visits (95% interval: 26.8–31.0). Restricting each model–schema comparison to questions with valid answers in both conditions gives gains of 37.4 and 8.4 points, and a direct difference of 29.0 (26.9–31.1). This selection-conditioned analysis does not recover a causal effect on model ability. Alternatively, treating absent records as unknown binary outcomes gives worst-case gain bounds of [33.2, 37.5] for endpoints and [4.2, 8.6] for visits, retaining available invalid answers as wrong. These are identification bounds, not confidence intervals. Strict scores remain primary. The accompanying script records all per-model denominators, uses 10,000 joint video replicates, and fixes seed 20260923.

B.3 COLOR CONTROLS

Across the unpaired synthetic subsets, multicolored-minus-monochrome overall strict accuracy ranges from +6.9 to +50.1 points across eleven selected model rows; across the ten available real rows it ranges from -2.2 to +14.9 points. These are descriptive differences between videos: synthetic color covaries with swap range, while the real levels vary color together with object counts and duration (Table 7).

Table 10 instead compares each marker with the unmarked rendering of the same multicolored synthetic video. These paired controls do not support the large gains seen on monochrome videos and remain separate from the confounded comparison between color subsets.

Table 10: Paired marker-minus-unmarked changes (pp) on the same multicolored synthetic videos at 2 fps, with 95% video-cluster intervals. Each available comparison contains 4,540 scheduled questions across 250 videos per model; no 397B colored-box result is available.

Model	Labeled box	Colored box	Colored point
Qwen3.5-4B	+0.4 [-1.5, +2.2]	-3.6 [-5.5, -1.6]	-6.5 [-8.2, -4.7]
Qwen3.5-9B	-4.1 [-5.9, -2.1]	-6.4 [-8.2, -4.5]	-13.9 [-15.8, -12.1]
Qwen3.5-27B	-0.6 [-1.9, +0.6]	-1.1 [-2.4, +0.1]	-18.5 [-20.5, -16.6]
Qwen3.5-35B-A3B	-0.5 [-2.1, +1.0]	-4.8 [-6.5, -3.1]	-17.8 [-19.8, -15.9]
Qwen3.5-122B-A10B	-2.3 [-3.7, -0.8]	-1.8 [-3.1, -0.3]	-13.9 [-15.9, -11.8]
Qwen3.5-397B-A17B	-0.4 [-1.6, +0.8]	—	-44.6 [-50.3, -38.7]
Gemma-4-31B-it	-1.9 [-4.0, +0.2]	-2.1 [-3.8, -0.5]	-7.2 [-9.4, -5.1]
InternVL3.5-38B	-1.0 [-2.8, +0.7]	-1.2 [-3.0, +0.6]	+1.2 [-0.7, +3.0]

Strict accuracy retains all scheduled questions, including missing or invalid answers. These within-video changes are distinct from monochrome–multicolored subset differences, which compare different videos.

B.4 SEQUENCE LENGTH AND FRAME SAMPLING

Table 11 gives the exact per-model estimates and support behind the swap-count comparison. The 30–50-swap group is a separate design level, and the dataset contains no 11–29-swap videos. Groups contain different videos and jointly vary in duration, input length, cup count, and question mix; they are not repeated measurements of error accumulation within one video.

Table 11: Exact values behind the matched ball-state horizon figure. Model cells are unmarked → labeled-box strict accuracy (%). Counts are per model; videos are the bootstrap clusters. The final bin has actual support 30–50 and there are no 11–29-swap videos.

Model	1–3	4–5	6–8	9–10	30–50
Qwen3.5-4B	38.1→59.7	33.1→49.7	27.1→39.8	25.2→32.6	15.5→18.5
Qwen3.5-9B	48.2→67.5	39.9→53.9	36.8→47.8	30.5→43.4	21.8→25.2
Qwen3.5-27B	70.9→91.4	51.6→81.9	41.6→75.8	39.4→71.4	25.8→37.8
Qwen3.5-35B-A3B	63.4→86.3	45.9→76.3	37.4→71.4	34.6→63.1	24.4→31.3
Qwen3.5-122B-A10B	64.8→88.6	47.7→78.7	40.9→74.2	38.3→68.1	0.0→33.8
Qwen3.5-397B-A17B	75.3→92.1	53.3→84.9	43.5→78.9	41.4→74.5	3.7→43.6
Gemma-4-31B-it	66.6→78.9	49.4→65.9	41.9→64.9	33.5→56.7	25.1→36.1
InternVL3.5-38B	47.0→49.5	36.7→38.1	28.8→31.7	26.3→28.7	20.3→22.5
Questions/model	1,378	1,160	1,540	930	1,600
Videos/model	98	65	94	63	80

Figure 6 shows the corresponding model-specific curves and uncertainty. Lines connect categorical group means rather than consecutive states of one video.

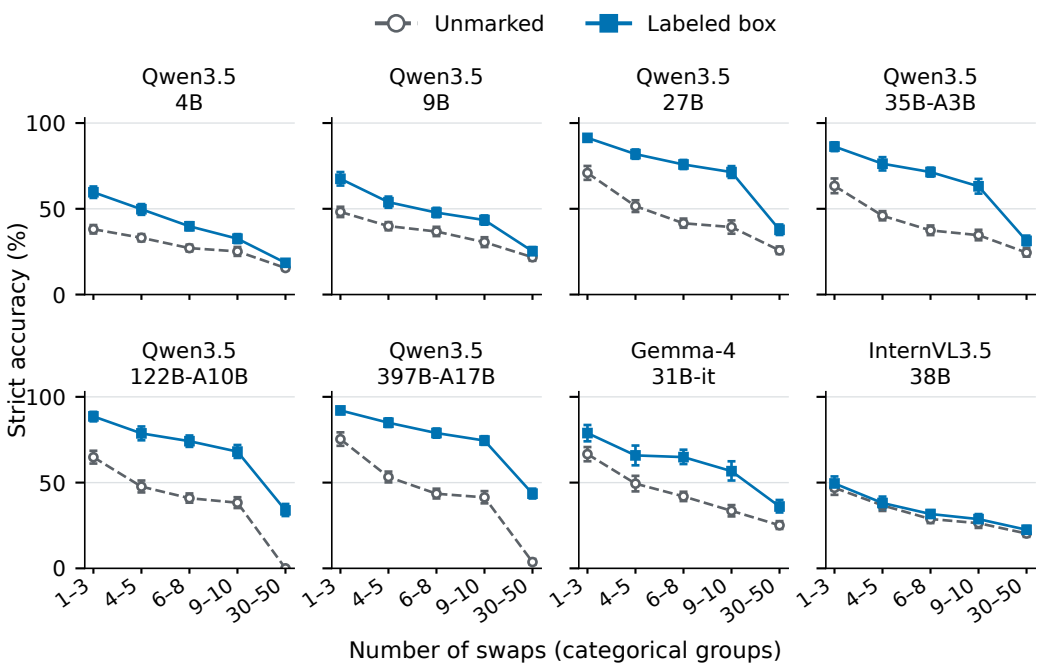


Figure 6: Matched ball-state strict accuracy at 2 FPS on monochrome videos. Error bars are 95% video-cluster intervals. The horizontal groups are categorical, with no 11–29-swap observations. All models use the per-bin question and video counts reported in Table 11.

Horizon sensitivity. The nine-cup, final-ball-position regression retains a negative slope after excluding the two models with incomplete baseline records: -0.69 points per swap without markers (95% interval: -0.92 to -0.52) and -1.21 with labels (-1.42 to -1.03), using 1,920 matched pairs from the same 70 videos. Conditioning all eight models on jointly valid answers instead gives -0.68 (-0.96 to -0.48) and -1.26 (-1.46 to -1.10). Both checks preserve the direction. The linear summary spans the unobserved 11–29-swap gap and does not control duration or input length.

Table 4 consolidates every available 1–4 FPS condition. Requested sampling rate can change both temporal evidence and input length; the sweep does not isolate these effects.

FPS validity sensitivity. At 4 versus 2 FPS, restricting to questions valid at both rates gives changes of $+1.4$ points for 122B (95% interval: $+0.5$ to $+2.3$; 11,588 questions), $+1.4$ for 9B ($+0.5$ to $+2.4$; 13,619), -2.5 for 4B (-3.5 to -1.6; 13,395), and -1.6 for Gemma (-2.6 to -0.7; 13,110). The 27B and 35B intervals still include zero. These selected populations differ from the scheduled-question denominator. If absent 122B records are instead assigned arbitrary binary correctness, its gain is bounded by $[-10.0, +5.0]$ points; a positive gain is therefore not identified without an assumption about missing outcomes. The mixed sampling pattern remains, while the large strict 122B gain is sensitive to record availability. Scripts retain the primary strict scores and export all six sensitivity estimates for models evaluated at every rate.

Figure 7 shows how the same FPS results vary across duration groups. Comparisons across FPS within a model and duration group keep the scheduled questions fixed, whereas duration groups contain different videos and question mixtures.

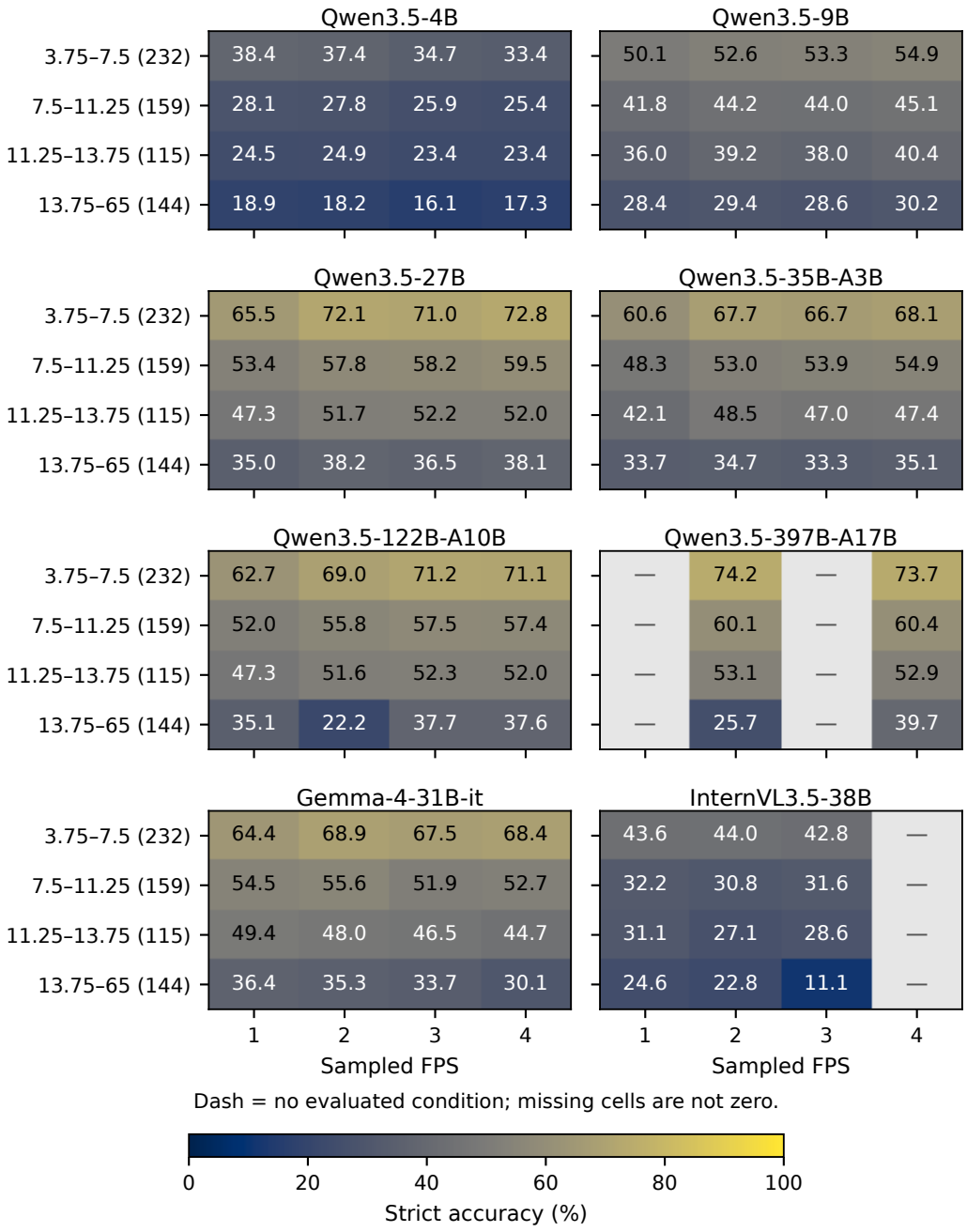


Figure 7: Strict accuracy (%) by video duration and requested uniform sampling rate for every available open-model condition. Duration bins are right-closed, with the first including the minimum; parentheses give unique-video support. All panels share a color scale. Gray cells are unavailable conditions, not zero accuracy.

B.5 RETROSPECTIVE TRAJECTORY RECONSTRUCTION

Figure 8 expands main Figure 4 by model. Eligible monochrome support falls from 400 videos at step 0 to 3 at step 50; multicolored support falls from 250 at step 0 to 38 at step 10.

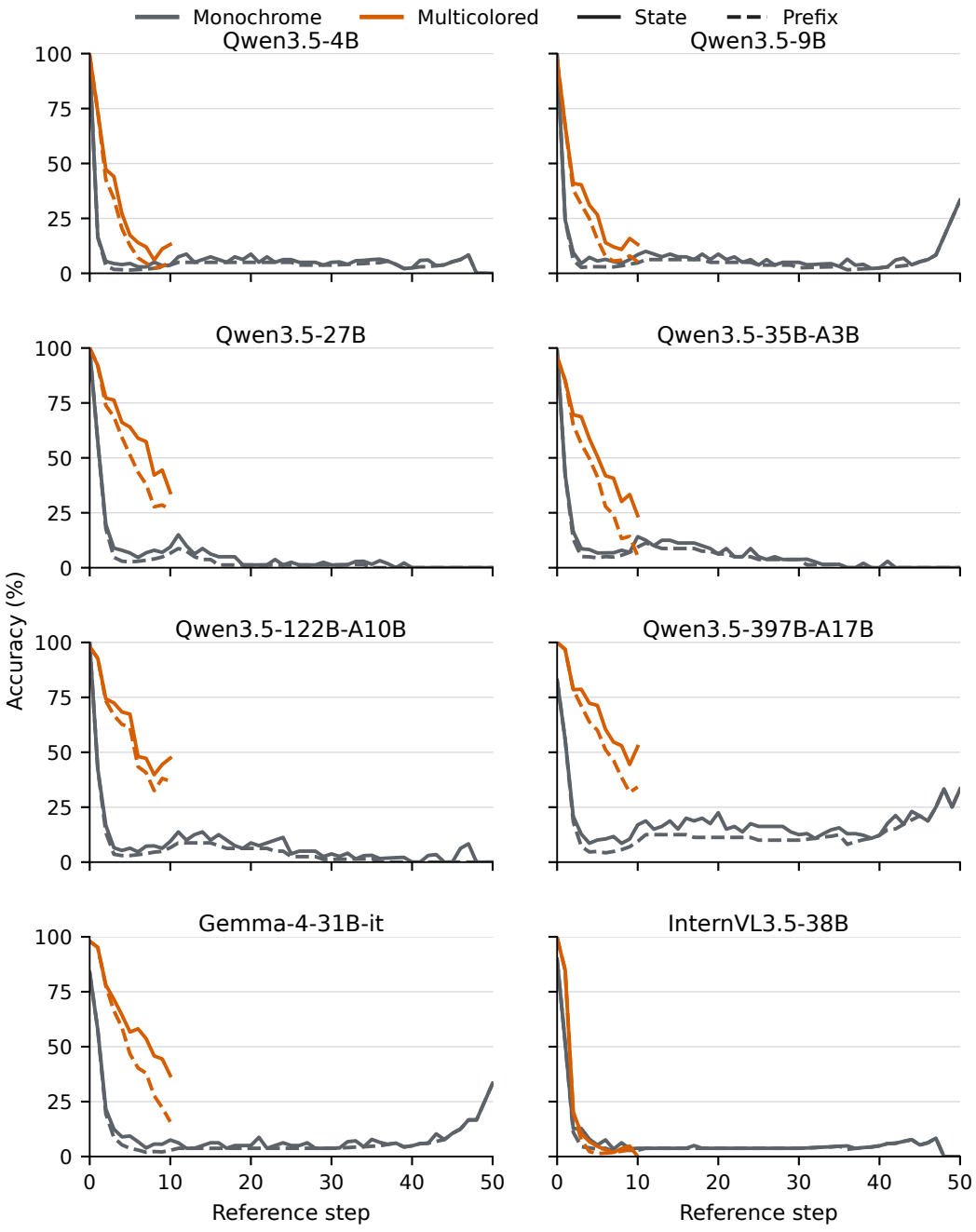


Figure 8: Retrospective trajectory reconstruction at 2 FPS: state accuracy (solid) and prefix correctness (dashed) for eight models. Color subsets contain different videos; later steps have changing support. Multicolored sequences end by step 10.

Figure 9 shows the joint cup-count/step pattern separately for each of the eight models, pooling the color subsets. Populated cells contain 1–120 unique videos, shared across models.

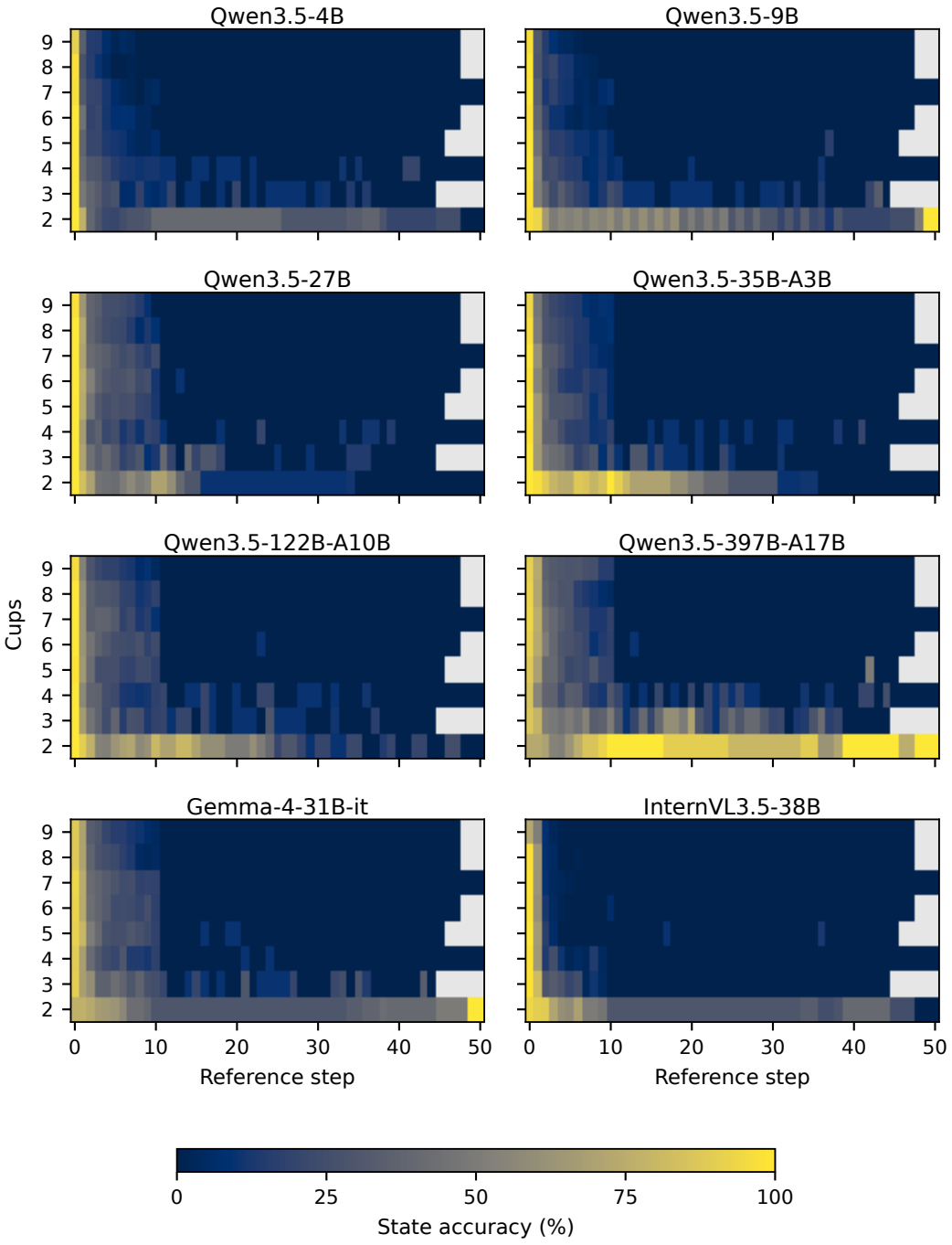


Figure 9: Per-state accuracy (%) jointly grouped by cup count and reference step for all eight models at 2 FPS. Monochrome and multicolored videos are pooled. Panels share a color scale; gray cells have no eligible video. Sparse late-step cells should not be read as precise estimates, and changing support prevents treating rows as repeated measurements of a fixed cohort.

C PROMPTS

The selected question and identifier follow the displayed user prefix. The optional R1-style system prefix specified in InternVL configurations is omitted because its use cannot be verified for every reported result.

In these prompts, a cup ID is the object's initial ordinal identity in the task logic (C1, C2, …); it is not necessarily a label drawn on the cup. Visual markers such as a bounding box with an ID or a colored point marker are separate interventions on the video. In the unmarked condition, IDs are defined only by the task and output conventions.

C.1 SYNTHETIC CUPS-AND-BALLS

System prompt for synthetic cups-and-balls

```
You are solving a deterministic state-tracking task.
We have a fixed number of cups arranged in a row from left to right.
Each cup has a unique identifier (ID), and positions are fixed.  Each
cup may have zero or more balls underneath it.  Cups can move, and
balls always move together with their cup.
DEFINITIONS:
- Cups are identified as:  C1, C2, C3, ..., CN (these identifiers
NEVER change).  - Positions are identified as:  P1, P2, P3, ...,
PN (from left to right).  - A state is a mapping from positions
to cups.  - Displacement = final_position - initial_position -
Negative displacement = movement to the left - Positive displacement
= movement to the right - If a question requires a yes/no answer,
answer "Yes"/"No" - The FIRST MOVEMENT is the first swap operation.
Lifting or revealing a cup (to show the ball) does NOT count as a
movement and is NOT a swap.
INITIALIZATION: You will receive the initial state as an ordered
list:  Example:  [C1, C2, C3] means:  P1 -> C1 P2 -> C2 P3 -> C3
OPERATIONS: Each action is a SWAP between two positions:  swap(Pi,
Pj) means:
- The cup at position Pi moves to Pj - The cup at position Pj moves
to Pi - Other cups are not affected
```

User prompt for synthetic cups-and-balls

```
RULES (STRICT):
1.  Cup identifiers NEVER change.  2.  Each position always contains
exactly one cup.  3.  A cup cannot occupy two positions at the same
time.  4.  The total number of cups is constant.  5.  A ball is
always associated with the cup covering it and moves together with
that cup.  6.  Positions without a ball are allowed.  7.  YOU MUST
UPDATE THE FULL STATE AFTER EVERY OPERATION. 8.  DO NOT SKIP STEPS.
9.  DO NOT GUESS THE ANSWER WITHOUT INTERMEDIATE STATES.
REASONING PROCEDURE: For each step:
- Start from the previous state - Apply the swap operation explicitly
OUTPUT FORMAT (STRICT JSON):
You will receive ONE question at a time together with its identifier
(for example, Q_3).  Return EXACTLY ONE JSON object with a single key
-- the identifier of that question:
{"Q_3":  <answer>}
- <answer> is a value of the type required by the question (see
below).  - Output nothing before the opening { and nothing after
the closing }.  - Do not wrap the JSON in ``` or add explanations.  -
Any reasoning must be fully completed before the opening {.
The answer format must exactly match the required answer type:
- Numeric answer:  Return only the number.  Example:  5
- Color answer:  Return only the color name.  Example:  red
```

```
- Object identifier: Return only the object identifier. Example:
C_3
- Ordered list: Return only a JSON array of identifiers in
left-to-right order. Example: ["C_1", "C_3", "C_2"]
- Mapping between objects: Return only a JSON object specifying the
associations between identifiers. Example: { "B_1": "C_2", "B_2":
"C_1" }
- Yes/no answer: Return only: "Yes" or "No"
- If an object cannot be identified from the video: Return only:
"undetermined"
IMPORTANT:
- Treat this as a state-tracking task, not a guessing task. -
Consistency is more important than speed. - Positions are numbered
from 1, from left to right. - Cup identifiers: C_1, C_2, ... (from
left to right in the initial state). - Ball identifiers: B_1,
B_2, ... (from left to right in the initial state). - Use only the
data type required by the specific question. - Do not turn numbers,
lists, or objects into strings. - Do not change the answer format.
- Do not add any text before or after the JSON object.
QUESTIONS (answer in the required format):
```

C.2 TRAJECTORY RECONSTRUCTION

In the trajectory-reconstruction protocol, the model must return the complete cup-state trajectory after every swap rather than answer an individual QA question.

System prompt for trajectory reconstruction

```
You are solving a deterministic state-tracking task.
We have a fixed number of cups arranged in a row from left to right.
Each cup has a unique identifier (ID), and positions are fixed. Each
cup may have zero or more balls underneath it. Cups can move, and
balls always move together with their cup.
DEFINITIONS:
- Cups are identified as: C1, C2, C3, ..., CN (these identifiers
NEVER change). - Positions are identified as: P1, P2, P3, ..., PN
(from left to right). - A state is a mapping from positions to cups.
- A state trajectory consists of the initial state and the full state
after every subsequent swap. - The FIRST MOVEMENT is the first swap
operation. Lifting or revealing a cup (to show the ball) does NOT
count as a movement and is NOT a swap.
INITIALIZATION: You will receive the initial state as an ordered
list: Example: [C1, C2, C3] means: P1 -> C1 P2 -> C2 P3 -> C3
The initial state is step 0 of the state trajectory.
OPERATIONS: Each action is a SWAP between two positions: swap(Pi,
Pj) means:
- The cup at position Pi moves to Pj - The cup at position Pj moves
to Pi - Other cups are not affected - After each swap, record the new
full left-to-right cup order - The first swap produces step 1, the
second swap produces step 2, and so on
```

User prompt for trajectory reconstruction

```
RULES (STRICT):
1. Cup identifiers NEVER change. 2. Each position always contains
exactly one cup. 3. A cup cannot occupy two positions at the same
time. 4. The total number of cups is constant. 5. A ball is
always associated with the cup covering it and moves together with
that cup. 6. Positions without a ball are allowed. 7. YOU MUST
```

```
UPDATE THE FULL STATE AFTER EVERY OPERATION. 8.  DO NOT SKIP STEPS.
9.  DO NOT GUESS THE FINAL STATE WITHOUT INTERMEDIATE STATES.
REASONING PROCEDURE: For each step:
- Start from the previous state - Apply the swap operation explicitly
OUTPUT FORMAT (STRICT JSON):
Return EXACTLY ONE JSON object in the following format:
{ "states":  [ { "step":  0, "cups":  ["C_1", "C_2", "C_3"] }, {
"step":  1, "cups":  ["C_2", "C_1", "C_3"] }, { "step":  2, "cups":
["C_2", "C_3", "C_1"] } ] }
- "states" is the full list of cup states in chronological order.
- "step":  0 is the initial cup order BEFORE the first swap.  -
Each subsequent "step" corresponds to the state IMMEDIATELY AFTER
the next swap.  - The "step" number increases by 1 after each swap.
- "cups" is the full left-to-right cup order at that step.  - The
number of elements in "states" must equal the number of swaps + 1.
- Lifting or revealing a cup does not count as a swap and does not
add a new state.  - If there are no swaps, return only the state
with "step":  0.  - Output only the sequence of cup states.  - Output
nothing before the opening { and nothing after the closing }.  - Do
not wrap the JSON in ``` or add explanations.  - Any reasoning must
be fully completed before the opening {.
IMPORTANT:
- Treat this as a state-tracking task, not a guessing task.  -
Consistency is more important than speed.  - Positions are numbered
from 1, from left to right.  - Cup identifiers:  C_1, C_2, ...  (from
left to right in the initial state).  - Ball identifiers:  B_1, B_2,
...  (from left to right in the initial state).  - Each state must
contain every cup exactly once.  - Do not change the answer format.
- Do not add any text before or after the JSON object.
```

C.3 REAL CUPS-AND-BALLS

System prompt for real cups-and-balls

```
You are solving a deterministic tracking task.  We have a fixed
number of cups placed in a row from left to right.  Each cup has a
unique identity (ID), and positions are fixed.  Each cup may have
zero or more balls under it.  Cups can move, and balls always move
together with their cup.
DEFINITIONS:
- Cups are labeled:  C1, C2, C3, ..., CN (these identities NEVER
change).  - Positions are labeled:  P1, P2, P3, ..., PN (from left
to right).  - The state is a mapping from positions to cups.  - Shift
= final_position - initial_position - Negative shift = moved left -
Positive shift = moved right - If the question requires a yes or no
answer, answer True/False without quotes - The FIRST MOVEMENT is the
first swap operation.  Lifting or revealing a cup (to show the ball)
is NOT a movement and does NOT count as a swap.
INITIALIZATION: You will be given the initial state as an ordered
list:  Example:  [C1, C2, C3] means:  P1 -> C1 P2 -> C2 P3 -> C3
OPERATIONS: Each action is a SWAP between two positions:  swap(Pi,
Pj) means:
- The cup at position Pi moves to Pj - The cup at position Pj moves
to Pi - No other cups are affected
```

User question prompt for real cups-and-balls

```
RULES (STRICT):
1.  Cup identities NEVER change.  2.  Each position always contains
exactly one cup.  3.  No cup can appear in two positions at once.  4.
```

```
The total number of cups is constant.  5.  You MUST update the full
state after EACH operation.  6.  Do NOT skip steps.  7.  Do NOT guess
the final answer without intermediate states.
REASONING PROCEDURE: For each step:
- Start from the previous state - Apply the swap operation explicitly
COLOR CONSTRAINTS:
If the answer requires a color value, use ONLY: [red, green, blue,
yellow, orange, purple, black, white, pink, brown]
OUTPUT FORMAT (STRICT JSON):
You will receive ONE question at a time together with its question ID
(for example, Q1).  Return EXACTLY ONE JSON object with a single key
-- the ID of that question:
{"Q1":  <answer>}
- <answer> is the value of the type required by that question (see
below).  - Output nothing before the opening { and nothing after
the closing }.  - Do not wrap the JSON in ``` and do not add any
explanation.  - Any reasoning must fully end before the opening {.
The answer format MUST match the required type of the question:
- Questions requiring a single numeric answer (Q1, Q2, Q3):  Return
only an integer number.  Example:  5
- Questions requiring a single boolean answer (yes/no):  Return only
True or False without quotes.  Example:  true
- Questions requiring a mapping between entities (Q4, Q5, Q6):
Return only a JSON object where keys are entity IDs and values are
the corresponding attributes.  Example:  { "B1":  "C3", "B2":  "C1" }
- Questions requiring an ordered sequence/list (Q7):  Return only
a JSON array preserving the required order.  Example:  ["C2", "C1",
"C3"]
- Questions requiring a list of items without ordering:  Return only
a JSON array.  Example:  ["B1", "B2", "B3"]
IMPORTANT:
- Treat this as a state-tracking problem, not a guessing task.  -
Consistency is more important than speed.  - The returned value must
exactly match the expected data type.  - Do not convert numbers,
booleans, objects, or arrays into strings.  - Do not add any
additional text before or after the JSON object.
QUESTIONS (return only answer in required format):
```