CHRONOSPATIAL: A PROCEDURAL VIDEO BENCHMARK FOR MEMORY AND WORLD-STATE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video questions about earlier events require a model to recover the relevant moment and distinguish past state from later changes. We introduce ChronoSpatial, a collection of 500 procedural videos with executable scene plans, four-choice puzzles, and deterministic answer programs. We evaluate seven open-weight video models and an isolated agent that selects timestamped views under explicit frame and call budgets. We report five evaluation cohorts separately, retaining development examples outside their headline scores. On the original 195-item evaluation, direct models achieve 25.1% to 35.9% accuracy. On the final 61 videos, they achieve 24.6% to 39.3%, while the Astra agent scores 68.9%, 93.4%, and 100% as its inspection allowances increase. Family-level results, matched input controls, state probes, and inspection traces show where the tested systems struggle and what information helps. Chronological video benefits one model in selected later-cohort comparisons, but this advantage does not hold across all tested models and tasks. An initial small fine-tuning pilot fails to improve held-out accuracy and exhibits answer-position bias. A label-balanced follow-up scores 18/56 on the reused test set, compared with 14/56 for the base model, but transfer falls from 11/34 to 9/34 and the test-change interval includes zero. We retain the construction plans to support stronger supervision and controlled follow-ups. The benchmark measures budgeted video inspection; its current results do not establish frontier failure or isolate internal memory capacity. Two human reviewers judged 50 randomly selected video–puzzle pairs solvable.

1 INTRODUCTION

A person leaves a book on a desk, moves it while preparing food, and later carries it past the television. A question can refer to the book's current location, its location during an earlier activity, or the object involved in a particular previous delivery. These questions share the same visual material but require different histories. Camera motion adds another source of variation: an object can change its position in the image while remaining at the same station in the room.

Video accuracy alone leaves several explanations open. A system may miss a short placement, confuse two objects, retrieve the wrong activity boundary, or apply the question incorrectly to a correct event history. An agent can also answer by repeatedly returning to the relevant frames. This behavior is useful in an inspection task, but a high score under that interface provides limited evidence about memory after a single viewing.

ChronoSpatial makes the scene history executable. A structured description defines entities, initial state, timestamped events, and a camera plan. A renderer produces the video, while a typed program computes the answer from the event history. The solver receives the video and question through a restricted interface. The private history supports scoring and the construction of diagnostic interventions. Figure 1 describes the implemented separation between scene construction and evaluation.

We contribute 500 rendered scenes with executable histories, a reproducible evaluation pipeline, and completed comparisons of visual, symbolic, and bounded-record inputs. We find large differences across task families and inspection allowances. The highest-budget agent solves nearly every evaluated scene, while Qwen3-VL-8B-Instruct scores only 5/12 when supplied with complete event

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

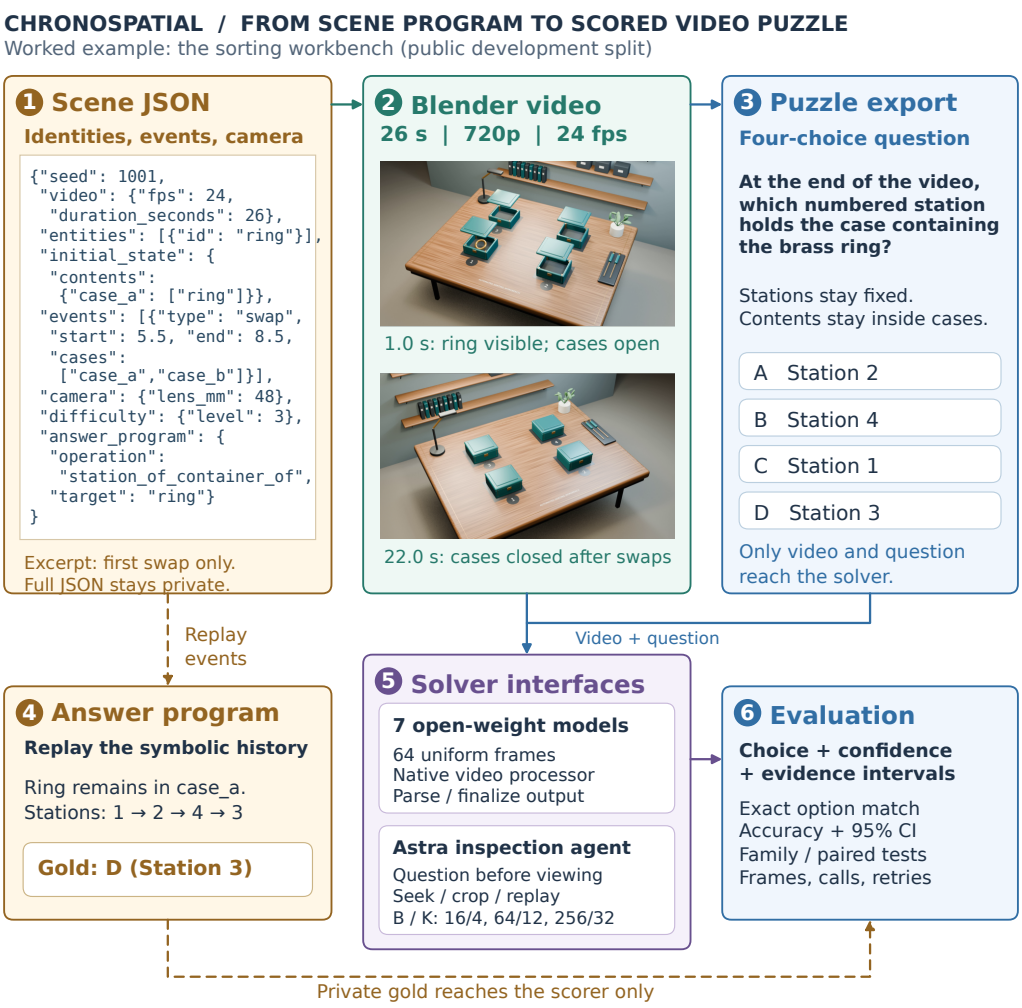


Figure 1: From executable scene descriptions to scored video puzzles, illustrated with a public development example excluded from the 195-item evaluation split. **(1) Scene specification.** The JSON excerpt shows the seed, video settings, persistent object identity, initial containment, one timestamped swap, camera parameters, difficulty level, and answer program; the complete scene contains the remaining events and rendering settings. **(2) Video construction.** Blender renders the scripted history. Actual frames at 1 and 22 seconds show the ring before lid closure and the final arrangement after case swaps and camera motion; the two thumbnails summarize a 26-second clip, not the solver's full observations. **(3) Puzzle export.** The video, question, and four choices enter the evaluation interface. Fixed station numbers require tracking the ring's case rather than its image position. **(4) Reference answer.** The controller replays all events: the ring stays in its case as that case visits stations $1 \to 2 \to 4 \to 3$, yielding option D. This development answer is shown only to explain scoring. **(5) Model inference.** Direct models receive 64 uniform frames through native processors, with output finalization when needed. The isolated Astra agent receives the question first and requests timestamped views under joint frame/call allowances (B, K), defined in Section 4; revisits and crops consume the allowance. **(6) Evaluation.** The controller matches the submitted choice to its private key and aggregates accuracy, uncertainty, family and paired-control comparisons, and inspection/finalization costs. Solid connectors carry construction outputs or solver-visible inputs; dashed connectors carry private state processing and gold answers, which bypass the solver.

transcripts in our diagnostic cohort. Our initial fine-tuning pilot exhibits answer-position collapse.

Balancing labels and rotating options raises accuracy on a reused test set, but does not improve family transfer. These results motivate intermediate-state supervision and evaluation on fresh histories.

2 Related work

Controlled video datasets connect actions to symbolic state: CATER studies object tracking, CLEVRER studies causal and counterfactual physical questions, and Physion compares physical prediction with human judgments (Girdhar & Ramanan, 2020; Yi et al., 2020; Bear et al., 2022). STAR and AGQA use structured representations for compositional questions (Wu et al., 2024a; Grunde-McLaughlin et al., 2021). ChronoSpatial builds on these ideas with executable histories and retrospective references; it does not claim that programmatic questions or symbolic scoring are new.

TemporalBench, TOMATO, and TempCompass motivate tests of frame order, single-frame shortcuts, and answer biases (Cai et al., 2024; Shangguan et al., 2025; Liu et al., 2024). EgoSchema distinguishes video length from the evidence needed to answer a question (Mangalam et al., 2023). We similarly separate chronological input, selective inspection, and bounded external records. MentisOculi motivates our comparisons with supplied symbolic and intermediate state information (Zeller et al., 2026). Its mental-imagery task differs from our retrieval of observed histories. Appendix A situates the benchmark among long-video, spatial, streaming, and physical-reasoning datasets.

3 Dataset construction

3.1 Scenes, histories, and questions

Let S_0 be the initial symbolic state and $e_{1:m}$ the ordered sequence of m events. After event e_j, the state is $S_j = F(S_{j-1}, e_j)$, where F is the deterministic transition rule. At time t, the renderer R produces frame $V_t = R(S(t), C_t, L_t)$ from the animated scene state $S(t)$, camera pose C_t, and lighting L_t. We compute the correct option y^* with an answer program g:

$$y^* = g(S_0, e_{1:m}, q, \mathcal{A}), \qquad y^* \in \{A, B, C, D\}. \tag{1}$$

Here V is the video, q contains the question and task instructions, and $\mathcal{A}$ is the ordered set of four labeled options. The solver receives these three inputs. We score its choice by exact matching to y^*.

The JSON scene representation includes persistent object identifiers, named locations, event times, question parameters, rendering settings, and a camera description. Depending on the family, state records containment, ownership, orientation, visit order, or object location. Videos use stylized procedural geometry and articulated people. Physics examples use stated analytic assumptions where applicable; the collection does not establish general physical fidelity.

The home/industrial history extension composes two references to history. A late activity boundary identifies a delivery, which identifies an object. The question then asks for an earlier location in that object's history relative to another activity. This construction prevents the final scene state from being a sufficient symbolic answer by design. Whether a rendered single frame or question wording supplies an unintended shortcut remains an empirical question for the controls.

Table 1 links the authoring plan to visible evidence and the scored question. We retain initial states, every timed event, camera settings, and answer programs in the accompanying `dataset_plans/` archive. These answer-bearing construction files are excluded from solver inputs.

3.2 Collection and splits

The original collection contains 100 initial scenes, 100 advanced scenes, five camera variants, and ten roofed home or industrial scenes. Its 16 family labels cover container tracking, board transformations, routes and landmarks, household errands, exchanges, queues, tools, and idealized physics. These labels reflect the implemented generators; they do not define 16 independent cognitive abilities. Durations range from 20 to 121 seconds, with a median of 49 seconds. Twenty development items support adapter work, leaving 195 original evaluation items. Five camera variants derive from four earlier scenes, so some observations share an underlying event history.

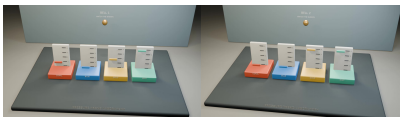
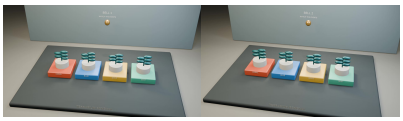
Construction plan	Video snapshots	Puzzle and answer
cs451: Four colored gauges change pressure on fixed 0–4 scales. The plan specifies every adjustment, two bell cues, and a moving camera.	 Bell 1: 21.5s; bell 2: 55.5s	Which gauge had the greatest net pressure increase from bell 1 to bell 2? A: Coral; B: Mint; C: Blue; D: Gold Answer: D (Gold)
cs471: Four pots receive replacement plants that reset plant age. Initially coral is oldest, then blue, gold, and mint. The plan fixes each replacement and two bell cues.	 Bell 1: 21.5s; bell 2: 55.5s	At bell 2, which terrarium held the oldest surviving plant? A: Blue; B: Mint; C: Gold; D: Coral Answer: D (Coral)

Table 1: Two examples from the final cohort. We show a short summary of each executable plan, decoded frames at the two bell cues, and the puzzle with its answer. We selected the first example in each named family without using model predictions. The frames illustrate the task; they are not a claim that these two images contain all necessary evidence. Complete JSON plans accompany all 500 videos.

The history extension contains 12 development and 12 validation videos in home and industrial settings. We call the validation cohort Challenge-12. Across its 24 rendered clips, durations range from 216 to 392 seconds, with a median of 350.5 seconds. Each scene tracks three visually distinct objects among four stations. Repeated activities, relocation histories, and camera motion increase the number of events a solver may need to inspect. All renders are 1280×720 at 24 frames per second. Each video has one primary four-choice puzzle; diagnostic questions on existing videos do not increase the dataset size.

Three later cohorts add 261 videos: Diverse-100, Mechanism-100, and Final-61. They cover 26 additional generator families, including accumulated quantities, logical circuits, cross-checkpoint comparisons, waiting-time priority, replacement age, and directed event transitions. We evaluate every item in these cohorts. Their durations are 46–74, 67–95, and 74–127 seconds, respectively. Together with the original 215 and the 24-video history extension, they bring the collection to 500. We keep the original 20 development and 12 extension development items separate: the headline cohorts contain 195, 12, 100, 100, and 61 items. These 468 items do not constitute one independent random sample. Appendix J gives the later construction and control protocols.

3.3 Construction checks and adaptive selection

We write and validate the scene plan before rendering it in Blender. For the extension, we check answers with two independently implemented symbolic procedures and test selected geometric conditions, including contact and camera visibility. These checks do not certify perceptual solvability. Two humans judged 50 randomly sampled pairs solvable in an offline review (Appendix G). Evidence intervals are annotated for only four original examples; general evidence-recall claims would therefore be premature.

The extension was developed after the original agent results were available. A fixed rule tested four development items at each of levels 2, 4, and 6, choosing the first level with accuracy at most 65%, or level 6 if none qualified. The levels jointly change history parameters and are not an independently controlled difficulty scale. We used the level-6 fallback; completed development scores were 4/4 at every level. We report the service retry in Appendix C. We evaluate all 12 level-6 validation items at three agent budgets, separately from development.

4 Evaluation protocol

4.1 Two interfaces

Direct models receive 64 uniformly sampled RGB frames through their native video processor, together with the question and options. Frames are resized to at most 448^2 pixels in area, preserv-

ing aspect ratio, before model-specific processing. The seven completed models are MiniCPM-V-4.6, GLM-4.6V-Flash, InternVL3.5-8B, Qwen3.5-9B, Qwen3-VL-8B-Instruct, Qwen3-VL-8B-Thinking, and Qwen3-VL-32B-Instruct. The 32B condition uses 4-bit NormalFloat (NF4) language weights and 16-bit brain floating-point (BF16) vision weights; the other six use BF16 throughout. We preserve checkpoint revisions, precision, sampler settings, and worker hashes in the run records. These baseline comparisons are inference-only experiments with deterministic decoding and seed 17.

The agent condition uses the service alias `gpt-6-astra`, with high reasoning effort, in a fresh isolated session for each item. It receives the question before inspection and can request timestamped frames through `probe` and `step`, query basic metadata, and submit an answer. The solver cannot inspect the generating code, answer key, parent conversation, or private state files. The image service supplies only the current video; unrestricted filesystem and network inspection tools are unavailable. The authenticated model transport is distinct from the restricted inspection service.

Agent observations are JPEG frames resized to at most 448^2 pixels in area, preserving aspect ratio. The joint allowances (B, K) are $(16, 4)$, $(64, 12)$, and $(256, 32)$, where B counts displayed frames and K counts tool calls, including submission. Repeated or cropped frames consume the allowance again. Sessions have a 600-second wall-clock limit. A service alias does not identify an immutable model checkpoint, and the service does not expose the same compute accounting as the local models.

These interfaces differ in frame selection, image encoding, reasoning implementation, and opportunity to revisit evidence. Equal nominal B therefore does not establish equal compute or equal visual input. We report each condition as a system result and use within-interface comparisons for paired statistical analyses.

4.2 Output validity and scoring

All systems in the inference benchmark return an option, a confidence value, and up to three evidence intervals. The separate adaptation pilot uses the single-letter contract in Section 6. The corrected direct protocol parses permitted presentation wrappers without consulting the answer key. When the output remains invalid, the same model receives its prior output and the same sampled frames for up to two constrained finalization attempts of 512 generated tokens each. The initial generation limit is 2,048 tokens. Each additional presentation and generation is recorded.

The completed corrected runs have no unresolved output errors. Finalization can still alter an answer and gives the model additional processing time. We therefore treat the corrected protocol as a complete inference condition. Its accuracy cannot be described as a reasoning gain caused by a parser fix. We retain the initial token-limit flags and finalization counts to make that distinction visible.

4.3 Matched input controls

For MiniCPM, InternVL, and Qwen3-VL-8B-Instruct, we evaluate question-only, middle-frame, and shuffled-frame conditions on the same fixed 64 evaluation items, four per original family (selection seed 260915502). Development smoke tests are separate. The reference is each model's uniform-video score on these exact items. Question-only supplies no images; middle-frame uses a single-image input path. Shuffling permutes the original 64 sampled frames and assigns chronological presentation timestamps, so timestamps do not disclose the original order. Model revisions, decoding settings, and finalization policy are retained. The controls change input length or modality, and finalization costs can differ.

For the three later cohorts, we retain the seven pinned models and three agent budgets. Controls use MiniCPM, Qwen3.5-9B, and Qwen3-VL-8B-Instruct, with question-only, final-frame, and shuffled-frame inputs on the same four items per family (40, 40, and 24 items). The final frame differs from the original cohort's middle-frame control. We compare each control with chronological video on its exact item subset.

Condition Videos	Original 195	Challenge 12	Diverse 100	Mechanism 100	Final 61
MiniCPM-V-4.6	56	3	26	29	15
GLM-4.6V-Flash	50	4	57	32	23
InternVL3.5-8B	50	5	33	23	24
Qwen3.5-9B	70	3	46	52	16
Qwen3-VL-8B-I	49	1	37	27	20
Qwen3-VL-8B-T	50	1	42	20	20
Qwen3-VL-32B NF4	58	4	33	27	16
Astra-16	154	4	73	79	42
Astra-64	189	11	100	99	57
Astra-256	190	12	100	99	61

Table 2: Correct answers in each separately evaluated cohort; denominators appear in the second header row. Direct models receive 64 uniform frames. Astra labels give the frame allowance; call limits are 4, 12, and 32. I/T denote Instruct/Thinking; 32B uses NF4 language weights. The Challenge-16 condition was added in a follow-up on the same twelve videos. Development examples are excluded.

4.4 Metrics and uncertainty

We report item accuracy and an unweighted mean of accuracies over the 16 original families. Four-way uniform guessing has expected accuracy 25%; the original evaluation answer counts are 52, 50, 45, and 48 for A through D. Challenge-12 has three answers of each label. Wilson 95% intervals summarize uncertainty in a binomial item model. Shared generators and related scenes limit that approximation.

Within the original agent conditions, we report paired accuracy changes, percentile intervals from 10,000 item bootstrap samples, and exact McNemar tests with Holm correction across the three budget comparisons. Separate Holm families cover the 21 pairwise direct-model comparisons and the nine control-versus-uniform comparisons. A sensitivity analysis excludes the five camera derivatives. These are exploratory analyses of one recorded run per condition, not estimates of service-level stochastic variation. The much smaller fresh validation cohort requires separate reporting.

5 Results across five cohorts

Table 2 reports all completed primary evaluations. On the original cohort, seven direct models score 49–70/195 (25.1–35.9%). On Final-61, they score 15–24/61 (24.6–39.3%). No direct model leads every cohort: GLM has the highest score on Diverse-100, Qwen3.5 on Mechanism-100, and InternVL on Final-61. These point estimates describe different task mixtures. They do not isolate an effect of video length or prove a general model ranking. Appendix B retains the original intervals, paired comparisons, and controls.

What inspection budgets change. Astra scores 42/61, 57/61, and 61/61 on Final-61. Moving from $(16, 4)$ to $(64, 12)$ gains 17 answers and loses two: a 24.6-point paired improvement (95% bootstrap interval [13.1,37.7], Holm-adjusted $p = 0.00146$). The next budget gains four answers, but its exact paired test remains inconclusive ($p = 0.125$). More broadly, its highest-budget scores across the five cohorts are 190/195, 12/12, 100/100, 99/100, and 61/61. Our harder generators have not established a sustained failure regime for this agent.

Figure 2 shows why allowance and use need separate reporting. At the 256-frame allowance, Astra displays an average of 42.5, 63.0, and 99.5 frames on the three later cohorts. Near-perfect scores therefore do not imply that it inspected every permitted frame. Question-first access and revisits make this a budgeted inspection result.

Where models fail. Figure 3 shows that aggregate scores hide family differences. Qwen3.5 scores 5/5 on home-interior questions but 1/8 on container tracking. Astra-64 scores 5/8 on 3D transformations despite 189/195 overall. These unequal, small denominators require caution. The 26-family breakdown in Figure 7 reveals substantial variation within cohorts. On Final-61 pressure trends,

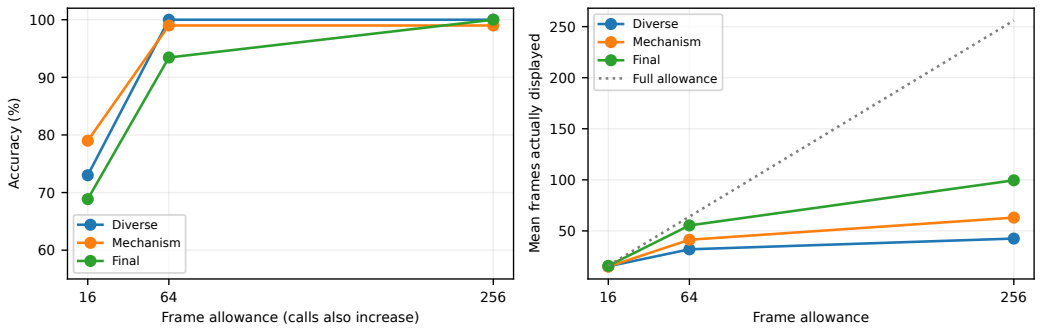


Figure 2: Later-cohort agent accuracy and actual inspection use. Frame and call allowances increase together. The right panel counts every displayed frame, including repeat views; the dotted line shows the full allowance. Each point is a separate run, not an extension of a previous run.

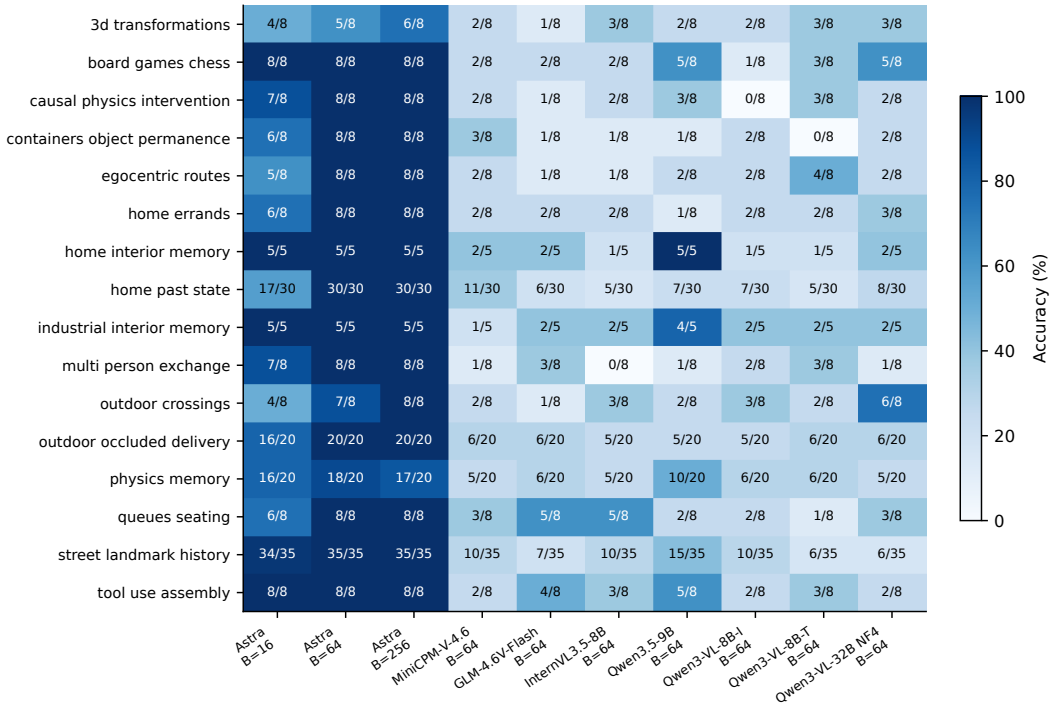


Figure 3: Completed original conditions by family. Cells give correct/total counts, including the small five-item interior families. The color scale is accuracy. These descriptive cells do not identify a causal effect of scene type: family labels combine different event structures and visual settings.

Qwen3.5 and the 32B model each score 0/10, while InternVL scores 6/10. On priority scheduling, GLM scores 8/10 and Qwen3-VL-8B-Instruct scores 1/10. Astra's four 64-frame errors concern two priority items, one transition-pattern item, and one pulley item. These small cells localize observed errors without establishing internal causes. Across the later cohorts, all seven direct models miss the same 11, 20, and six items; Astra-64 answers 11, 19, and five of those correctly. Thus the open-model failures alone do not demonstrate that the rendered questions are unanswerable.

Does chronological video help? The original controls show no observed advantage for chronological video, with all nine paired adjusted p-values at least 0.691. Later results are more selective. Qwen3.5 scores 21/40 with video versus 8/40 question-only on Diverse-100, a 32.5-point advantage (Holm-adjusted $p = 0.0398$). On Mechanism-100 it scores 21/40 with video versus 2/40 from the

Arm	Development Video	Development No video	Test Video	Transfer Video
Untouched base	4/24	8/24	14/56	11/34
Video-answer SFT	8/24	9/24	12/56	8/34
Question-only SFT	6/24	9/24	11/56	8/34

Table 3: Original 2B pilot, correct/total. Development selects checkpoints; test and transfer are excluded from adaptation. All arms use 32 frames when video is supplied. One seed.

Arm	Test	Transfer
Untouched base	14/56	11/34
Balanced video SFT	18/56	9/34
Balanced question SFT	12/56	9/34

Table 4: Balanced follow-up with video at test time, correct/total. The same test histories were evaluated previously. The full input controls appear in Table 15. One seed.

final frame, a 47.5-point advantage ($p = 3.43 \times 10^{-5}$). No Final-61 control comparison is significant after the same within-cohort correction. Table 16 reports every matched comparison. These input interventions change evidence, modality, and processing cost; they support a video advantage in selected configurations, not a pure measure of temporal memory.

6 Exploratory fine-tuning pilot

We fine-tuned Qwen3-VL-2B-Instruct on a separate retrospective split of 120 training, 24 development, 56 test, and 34 withheld-family videos. Exact history groups remain within partitions. Video-answer supervision is compared with the untouched model and a question-only training control. Both supervised fine-tuning (SFT) arms use rank-16 low-rank adaptation (LoRA) on language attention, a frozen visual encoder and projector, three epochs, effective batch size eight, learning rate 10^{-4}, and seed 17. Loss covers only the answer letter and termination token. Video inputs contain 32 uniform frames; all arms use the same constrained answer format. Development accuracy selects the checkpoint; Appendix I gives the full recipe and split qualifications.

Video supervision raises development accuracy from 4/24 to 8/24, but the same adapter answers 9/24 correctly without video (Table 3). Its test accuracy is 12/56 versus 14/56 for the base, and family-transfer accuracy is 8/34 versus 11/34. The paired test change is -3.6 percentage points, with a history-cluster bootstrap 95% interval of $[-20.0, 14.0]$. The adapter chooses B on 89/90 held-out video questions; B is the training majority (47/120). This pilot does not establish held-out training utility. We next test balanced labels and option rotation against a question-only training control; intermediate-state supervision remains untested.

Balanced-label follow-up. We repeated the two training arms with 30 examples per correct position and cyclic option rotation across epochs, retaining the split and optimization recipe. This post-hoc follow-up reuses previously evaluated test histories. Balanced video training improves test accuracy from 14/56 to 18/56; it scores 9/34 on family transfer and 17/56 without video. Balanced question-only training scores 12/56 with video. The video-minus-base test change is $+7.1$ points (history-cluster 95% interval $[-9.3, 24.1]$). Table 15 gives all input controls. One seed and a reused test set limit generalization; this experiment does not establish that option balancing alone produces a reliable visual-reasoning gain. Table 4 places the balanced follow-up beside the original pilot; its test gain does not extend to the withheld families.

7 Information and state diagnostics

We adapted MentisOculi's supplied-representation comparisons to 12 history-extension videos with new frozen questions. Astra answers all 12 correctly with a 128-token external record, a 512-token record, a full transcript, or complete state histories; it scores 1/12 without a record. The answering sessions for bounded records receive no images. This shows that the records preserve useful information on these examples, while ceiling scores leave differences between record sizes unresolved. Figure 4 and Tables 5 and 13 report the completed comparisons and interface-specific diagnostics.

We also ask Qwen3-VL-8B-Instruct to report the three object locations after each of four partial histories. These fresh-session prefix probes recover 51/144 object locations and only 1/48 complete three-object states. These are errors in explicit outputs, not direct measurements of latent memory. A separate generation-budget check reduces initial truncations from 13/32 to 5/32 when the limit

Condition	Qwen3-VL-8B-I	Astra
Complete event transcript	5/12	12/12
Complete state history	3/12	12/12
Question before viewing	4/12	12/12
Question after viewing	2/12	12/12
128-token record	2/12	12/12
512-token record	3/12	12/12
No record / question only	1/12	1/12

Table 5: Shared diagnostics on 12 histories with new frozen questions, correct/total. Interfaces differ in processing and compute. A separate native question-only repeat also scores 1/12. These scores do not isolate perception or memory errors.

rises from 2,048 to 8,192 tokens, with accuracy changing from 10/32 to 11/32. We report methods and denominators in Appendix H; strict scene counterfactuals and activation interventions remain future work.

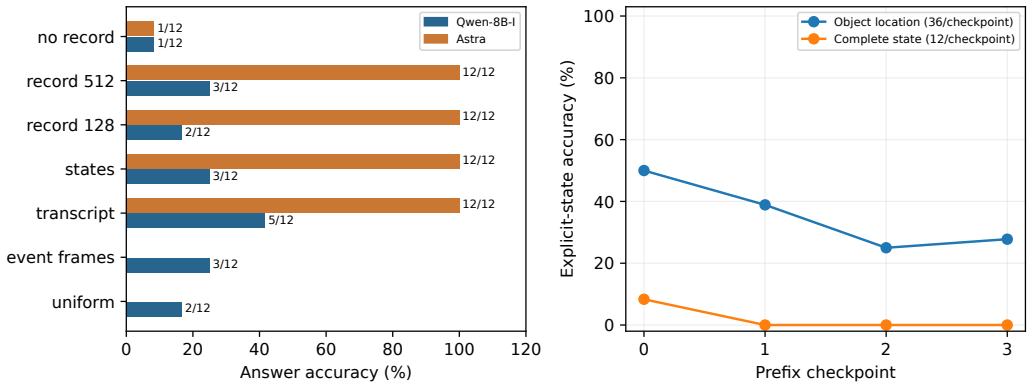


Figure 4: Completed information and state-output diagnostics. Left: selected frozen input conditions, all on the same 12 follow-up histories. Missing bars indicate unrun model/condition combinations. Uniform/event frames apply to the native model; symbolic and record conditions change representation as well as information access. Right: Qwen3-VL-8B-Instruct's fresh prefix probes, with 36 object locations and 12 full states per checkpoint. These explicit-output measures do not reveal internal representations.

8 Limitations and Next Steps

We evaluate stylized, procedurally related scenes with named stations and scripted behavior. Shared generators, sequential cohort design, and the original split's repeated histories limit independence. Our 50-item qualitative human review does not measure human accuracy or validate all 500 items.

We keep cohorts and interfaces separate because they differ in task mixture, inspection access, input processing, and output cost. The agent uses a service alias rather than an immutable checkpoint; the 32B baseline uses lower-precision language weights. Public availability also prevents a claim of training decontamination. Most results reflect one run, and the fine-tuning pilots do not estimate seed variability.

The balanced pilot leaves substantial answer-position bias. A next step is to use the construction plans to supervise object identity and intermediate state updates, then compare video-trained adapters with question-only controls on fresh histories. Human-reviewed decisive-event pairs and interventions on model-produced ledgers could test specific error explanations. Our current results do not identify an internal memory mechanism.

AI-USE STATEMENT

Generative AI assisted with research planning, hypotheses and experimental design, synthetic scene descriptions, Blender and evaluation code, debugging, result analysis, literature search, figure code, and manuscript drafting and editing. A tool-using AI system was also evaluated as a benchmark participant in separate isolated sessions. Deterministic programs compute answer keys and score submitted options. AI outputs are not counted as human ratings. The authors are responsible for checking the manuscript, references, and release artifacts. Two human reviewers conducted the offline solvability check described in Appendix G.

ETHICS STATEMENT

The videos depict procedural scenes and stylized people. The intended use is evaluation of visual state reasoning. We report an offline qualitative review by two humans on 50 randomly selected video–puzzle pairs. A future quantitative participant study should document consent, compensation, privacy, and the applicable institutional review determination. Performance on these scenes should not be used to certify safety in physical environments. Asset provenance and redistribution licenses require a completed inventory before dataset release.

REPRODUCIBILITY STATEMENT

We provide frozen aggregate result snapshots, source hashes, CPU analysis code, generated tables and figures, and all 500 answer-bearing JSON construction plans. These plans support reconstruction and supervision and are excluded from solver inputs. Experiment records retain model revisions, prompts, decoding settings, and inspection traces. The agent's service alias may change over time. Private answer keys and held-out scenes are kept separate from solver inputs. Appendix K publishes selected examples and answers after the result freeze for documentation; these pages are excluded from solver inputs. The manuscript source archive contains the plans and selected decoded frames; a complete anonymous video and evaluation-environment release still requires an asset-license audit. Future experiments are described in Appendix H and are not included in reported accuracy.

REFERENCES

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines, 2022. URL https://arxiv.org/abs/2106.08261.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models, 2024. URL https://arxiv.org/abs/2410.10818.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xiang An, Bo Li, Xin Xie, ZiDong Wang, Mingze Sun, Shuang Chen, Hongyu Li, Xiaobin Hu, and Ruqi Huang. 4dthinker: Thinking with 4d imagery for dynamic spatial understanding, 2026. URL https://arxiv.org/abs/2605.05997.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245.

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning, 2020. URL https://arxiv.org/abs/1910.04744. arXiv: 1910.04744.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense . In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5843–5851, Los Alamitos, CA, USA, October 2017. IEEE Computer Society. doi: 10.1109/ICCV.2017.622. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2017.622.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022. doi: 10.1109/CVPR52688.2022.01842.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning, 2021. URL https://arxiv.org/abs/2103.16002.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024. doi: 10.1109/CVPR52733.2024.02095.

Shaoxuan Li, Zhixuan Zhao, Hanze Deng, Zirun Ma, Shulin Tian, Zuyan Liu, Yushi Hu, Haoning Wu, Yuhao Dong, Benlin Liu, Ziwei Liu, and Ranjay Krishna. Perceptioncomp: A video benchmark for complex perception-centric reasoning. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *Computer Vision – ECCV 2026*, pp. 241–257, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-37553-7.

Junming Lin, Zheng Fang, Zihao Wan, Fuwen Luo, Chi Chen, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for MLLMs to achieve streaming video understanding, 2025. URL https://openreview.net/forum?id=qnAZqlMGTB.

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/.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y. Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning?, 2026. URL https://arxiv.org/abs/2505.23359.

Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, 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, 2023. URL https://arxiv.org/abs/2305.13786.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models, 2025. URL https://arxiv.org/abs/2410.23266.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Joshua B. Tenenbaum, Daniel LK Yamins, Judith E Fan, and Kevin A. Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=5Exz7eaBXH.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2025. URL https://arxiv.org/abs/2406.08035.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–442, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.25. URL https://aclanthology.org/2024.acl-long.25/.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos, 2024a. URL https://arxiv.org/abs/2405.09711.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: a benchmark for long-context interleaved video-language understanding. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024b. Curran Associates Inc. ISBN 9798331314385.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781, 2021. doi: 10.1109/CVPR46437.2021.00965.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2025. URL https://arxiv.org/abs/2412.14171. arXiv: 2412.14171.

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.

Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, and Jun Yu. Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23191–23200, 2023. doi: 10.1109/CVPR52729.2023.02221.

Jana Ricarda Zeller, Thaddäus Wiedemer, Fanfei Li, Thomas Klein, Prasanna Mayilvahanan, Matthias Bethge, Felix A. Wichmann, Ryan Cotterell, and Wieland Brendel. Mentisoculi: Revealing the limits of reasoning with mental imagery. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sxvuK2x3eA.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos, 2024. URL https://arxiv.org/abs/2410.02763.

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

A Extended related work

Controlled actions and physical state. Something-Something emphasizes fine-grained everyday interactions through action templates (Goyal et al., 2017). CATER uses compositional actions and object tracking in synthetic scenes, while CLEVRER asks descriptive, explanatory, predictive, and counterfactual questions about collisions (Girdhar & Ramanan, 2020; Yi et al., 2020). Physion compares visual physical prediction with human judgments and models given physical state; Physion++ adds inference about latent physical properties (Bear et al., 2022; Tung et al., 2023). These works establish controlled dynamics as a basis for diagnosis. ChronoSpatial focuses on historical identity and location across relocations, activity boundaries, and viewpoint changes. Its analytic physics examples do not match the breadth of physical phenomena or human validation in Physion.

Compositional video question answering. NExT-QA distinguishes causal and temporal action reasoning from description (Xiao et al., 2021). STAR constructs questions from situation hypergraphs and functional programs; AGQA studies indirect references and compositional generalization using scene-graph question programs (Wu et al., 2024a; Grunde-McLaughlin et al., 2021). ANetQA extends this approach to fine-grained semantics in untrimmed videos and tests auxiliary annotations and oracle frames (Yu et al., 2023). VideoReasonBench is especially close to our task because its operations update a partially visible state before recall, inference, or prediction questions (Liu et al., 2026). Our renderer and answer interpreter share an executable history, supporting intervention design. Programmatic question generation and symbolic answers are established techniques, and our separate adaptation pilot tests transfer to families withheld from its training split (Section 6).

Whether temporal evidence is necessary. MVBench assembles temporal tasks from existing video annotations; TempCompass tests changes in video dynamics (Li et al., 2024; Liu et al., 2024). TOMATO explicitly measures multi-frame gain, frame-order sensitivity, and frame-information disparity (Shangguan et al., 2025). TemporalBench examines fine-grained dynamics and answer-choice biases, while Vinoground uses short video-caption pairs with temporal counterfactual differences (Cai et al., 2024; Zhang et al., 2024). These studies motivate our question-only, single-frame, and shuffled-frame controls. Our original controls show no observed video advantage; later cohorts provide selected positive comparisons (Section 5). Scene complexity alone cannot establish temporal state tracking.

Memory and long-context understanding. Ego4D's episodic-memory tasks retrieve evidence about past first-person experience (Grauman et al., 2022). EgoSchema distinguishes clip duration from the temporal evidence needed to answer a question through temporal certificate sets (Mangalam et al., 2023). Perception Test includes memory and physics among its skill categories, and Mementos studies object and behavior errors across image sequences (Pătrăucean et al., 2023; Wang et al., 2024). Video-MME, MLVU, and LVBench broaden evaluation across domains, durations, and tasks; LongVideoBench uses referring questions over interleaved video and subtitles (Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025; Wu et al., 2024b). Our longest histories are a few minutes and are not an hour-scale context benchmark. We control event histories, but have not established general minimal evidence certificates or independent human accuracy.

Inspection, spatial reasoning, and diagnosis. StreamingBench asks questions as video unfolds, whereas PerceptionComp examines complex perception requiring repeated evidence gathering (Lin et al., 2025; Li et al., 2026). VSI-Bench studies spatial understanding and recall; 4DThinker develops dynamic spatial representations (Yang et al., 2025; Chen et al., 2026). These distinctions matter for our agent, which sees the question before inspecting frames and may revisit earlier timestamps. Its accuracy measures budgeted inspection, not retention after one viewing. MentisOculi's comparisons with symbolic descriptions and oracle intermediate images motivate our event-transcript and state-checkpoint conditions (Zeller et al., 2026). Like ANetQA's oracle-frame analysis, they test the benefit of supplied information (Appendix H), without identifying an internal memory mechanism.

B Original evaluation details

Table 6 gives the completed original results. The seven direct baselines achieve 25.1% to 35.9% accuracy in the evaluated configuration. Qwen3.5-9B has the highest point estimate, but none of

Model	Interface	B	Correct	Acc.	95% CI	Macro
Astra	agent	16	154/195	79.0	[72.7, 84.1]	79.8
Astra	agent	64	189/195	96.9	[93.5, 98.6]	96.2
Astra	agent	256	190/195	97.4	[94.1, 98.9]	97.5
MiniCPM-V-4.6	direct	64	56/195	28.7	[22.8, 35.4]	27.7
GLM-4.6V-Flash	direct	64	50/195	25.6	[20.0, 32.2]	27.7
InternVL3.5-8B	direct	64	50/195	25.6	[20.0, 32.2]	26.9
Qwen3.5-9B	direct	64	70/195	35.9	[29.5, 42.8]	38.8
Qwen3-VL-8B-I	direct	64	49/195	25.1	[19.6, 31.7]	24.5
Qwen3-VL-8B-T	direct	64	50/195	25.6	[20.0, 32.2]	28.4
Qwen3-VL-32B NF4	direct	64	58/195	29.7	[23.8, 36.5]	33.8

Table 6: Completed original evaluation conditions ($n = 195$). Accuracy, interval bounds, and family-macro accuracy are percentages. B is the nominal frame allowance. Direct models use the corrected finalization protocol; agent rows use different call allowances as well as different frame allowances. These are not equal-compute comparisons. I and T denote Instruct and Thinking; the 32B row uses NF4 language weights. Frozen results: September 18, 2026.

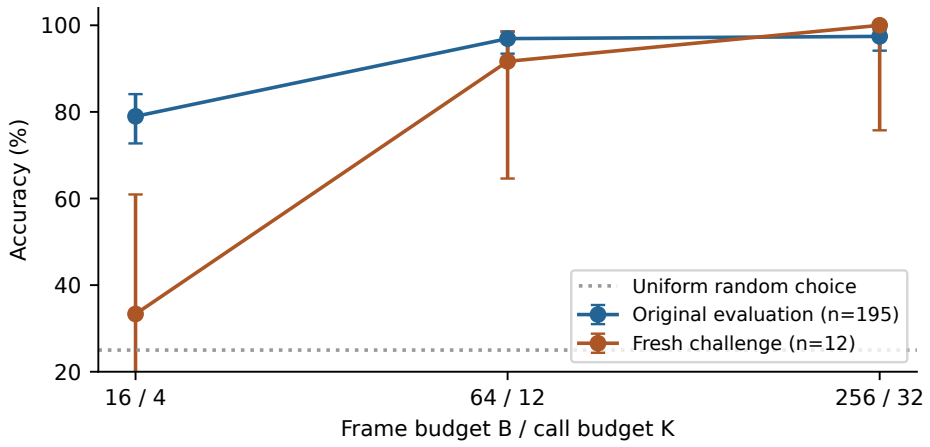


Figure 5: Agent accuracy under joint frame and call allowances. Error bars are Wilson 95% intervals. Each condition is a fresh recorded run; the plotted line does not imply that one run extends another. The 12-item challenge estimate is considerably less precise than the original estimate.

the 21 paired direct-model comparisons is significant after Holm correction (minimum adjusted $p = 0.324$). Its substantial truncation rate, discussed below, also limits conclusions about performance with a larger generation budget. The 32B NF4 condition achieves 58/195 (29.7%); its score does not establish a benefit from model size across different precision and implementation settings.

Figure 5 plots the original and challenge budget results. The agent improves from 154/195 at $(16, 4)$ to 189/195 at $(64, 12)$. This is a gain of 17.9 percentage points, with a paired bootstrap interval of 12.3 to 23.6 points. There are 37 newly correct items and two newly incorrect items; the Holm-adjusted exact McNemar p-value is 5.7×10^{-9}. Excluding camera derivatives gives an 18.4-point change on 190 items. Increasing the allowance again yields two gains and one loss, for a net change of 0.5 points and a bootstrap interval of -1.0 to 2.6 points. The frame and call allowances change together, so these observations do not isolate the effect of frame count.

The fresh extension does not reach the intended frontier failure regime (Table 7). Astra answers 4/12 correctly at 16 frames, 11/12 at 64 frames, and 12/12 at 256 frames. We added the 16-frame condition on September 24 using the same twelve videos and frozen inspection protocol; it was not part of the original result freeze. For the 64- and 256-frame conditions, the Wilson intervals are 64.6% to 98.5% and 75.8% to 100%, respectively. The seven direct models answer between 1/12 and 5/12 correctly; InternVL3.5-8B has the highest point estimate on this small cohort. We cannot

Model	Interface	B	Correct	Acc.	95% CI
Astra	agent	64	11/12	91.7	[64.6, 98.5]
Astra	agent	256	12/12	100.0	[75.8, 100.0]
MiniCPM-V-4.6	direct	64	3/12	25.0	[8.9, 53.2]
GLM-4.6V-Flash	direct	64	4/12	33.3	[13.8, 60.9]
InternVL3.5-8B	direct	64	5/12	41.7	[19.3, 68.0]
Qwen3.5-9B	direct	64	3/12	25.0	[8.9, 53.2]
Qwen3-VL-8B-I	direct	64	1/12	8.3	[1.5, 35.4]
Qwen3-VL-8B-T	direct	64	1/12	8.3	[1.5, 35.4]
Qwen3-VL-32B NF4	direct	64	4/12	33.3	[13.8, 60.9]
Astra	agent	16	4/12	33.3	[13.8, 60.9]

Table 7: Completed results on 12 fresh challenge validation videos. Development outcomes are excluded. The 100% point estimate remains compatible with a substantially lower success probability because of the small cohort.

Model	Uniform (64)	Question only (0)	Middle (1)	Shuffled (64)
MiniCPM-V-4.6	14/64 (21.9)	15/64 (23.4)	14/64 (21.9)	17/64 (26.6)
InternVL3.5-8B	16/64 (25.0)	18/64 (28.1)	19/64 (29.7)	19/64 (29.7)
Qwen3-VL-8B-I	11/64 (17.2)	16/64 (25.0)	19/64 (29.7)	12/64 (18.8)

Table 8: Matched controls on the same 64 items per model. Cells show correct/64 (percentage); header parentheses give the number of input frames. Uniform results are restricted to this subset and differ from the 195-item headline scores.

infer from this small, selected extension that longer histories are generally easy, nor can we attribute differences between its levels to a single difficulty factor.

B.1 Do the direct models benefit from chronological video?

None of the three tested models has a higher point estimate with chronological uniform video than with any of its controls (Table 8). Shuffling changes accuracy by $+1.6$ to $+4.7$ percentage points. The largest observed change is Qwen3-VL-8B-Instruct's single-frame result, 19/64 versus 11/64 with uniform video. All nine exact paired tests have Holm-adjusted $p \geq 0.691$; Appendix E gives differences and bootstrap intervals. These results do not establish equivalence between inputs or prove that a model ignores temporal evidence. Low baseline accuracy and the small cohort limit sensitivity.

The subset's answer counts are 14, 14, 12, and 24 for A through D. Uniform guessing still has expectation 25%, but always choosing D yields 37.5%. All observed control and matched reference scores fall below that majority-label baseline. The completed controls therefore provide no positive evidence that these configurations solve the questions through chronological state tracking. They leave open whether errors arise from observation selection, visual interpretation, generation limits, or state reasoning.

B.2 What the inspection traces show

At allowance 64, the agent displays a median of 42 frames on the original collection; at allowance 256, the median is 47. Thus, the larger allowance is often unused. On fresh validation, the medians increase to 64 and 125.5 frames. A temporal revisit means that at least one decoded timestamp occurs in more than one inspection; it can involve a different crop. Backward inspection means that the next successful image request begins before the latest timestamp in the preceding request. This includes returning from an overview to a local segment.

We also compute a geometric coverage proxy on the fresh validation clips. A relocation is covered if a viewed timestamp falls between the start of placement and the end of its settling interval. Averaged over items, the agent covers 35.8% of these windows at allowance 64 and 64.5% at allowance 256. A hypothetical uniform sampler using the full allowances covers 32.9% and 100%, respectively. This calculation is about timestamps only; it is not a model evaluation under uniform sampling.

Cohort	B	Median frames	Calls	Revisited time	Backward inspection
Original	16	16	3	28/195	177/195
Original	64	42	4	117/195	179/195
Original	256	47	4	120/195	144/195
Validation	64	64	6	2/12	12/12
Validation	256	125.5	5	9/12	10/12
Validation	16	16	3	0/12	11/12

Table 9: Recorded agent behavior. Medians count displayed frames and tool calls; the last two columns count items with at least one qualifying event. Temporal revisiting is external evidence access and does not reveal what the model retained internally.

Model	Finalization used	Initial truncation	Mean presentations	Unresolved
MiniCPM-V-4.6	195/195	5/195	2.02	0
GLM-4.6V-Flash	22/195	6/195	1.11	0
InternVL3.5-8B	13/195	0/195	1.07	0
Qwen3.5-9B	178/195	150/195	1.91	0
Qwen3-VL-8B-I	19/195	7/195	1.10	0
Qwen3-VL-8B-T	86/195	82/195	1.44	0
Qwen3-VL-32B NF4	0/195	0/195	1.00	0

Table 10: Costs of the corrected direct protocol on the original evaluation items. "Mean presentations" counts repeated presentation of the same 64 sampled frames. Initial truncation is the recorded generation-limit flag. Finalization and truncation are different events.

The agent can recover locations from other moments, and many relocations may be irrelevant to a particular answer. High answer accuracy with incomplete window coverage therefore does not establish a shortcut or faithful reconstruction of the full history.

B.3 Generation limits and finalization costs

Table 9 summarizes observed inspection costs and the placement-window coverage proxy. Table 10 reports a second source of variation. MiniCPM uses finalization on all 195 evaluation items. Qwen3.5 reaches the initial generation limit on 150 items and uses finalization on 178; Qwen3-VL-8B-Thinking reaches the limit on 82 items and uses finalization on 86. The 32B NF4 condition needs no finalization. These counts matter when interpreting the direct scores: a strict output contract and a finite reasoning budget are part of the measured system. Resolving all formatting failures removes missing answers from the final table, but does not show that the generation budget was sufficient for each model. A 32-item matched follow-up reduces truncation but changes accuracy by only one answer (Appendix H).

C Dataset inventory and documentation

Table 11 lists the original family counts. The extension adds 12 home-history and 12 industrial-history videos, half for development and half for fresh validation. These are new generator families with compound references to activity and object history. This initial collection has 239 rendered videos and one primary four-choice question per video; the later 261 videos are reported separately in Appendix J. Development examples are excluded from evaluation accuracy.

Intended use. The collection supports controlled study of historical state questions and evidence access. It does not sample the distribution of everyday human behavior, real-world navigation, or industrial safety tasks. Text descriptions and task instructions are in English. Visual appearance, spatial layouts, and scripted actors restrict the domain.

Provenance and access. Procedural JSON descriptions drive Blender construction. The internal build store contains events, answer programs, validation outputs, and seeds. Evaluation exports contain salted identifiers, videos, and questions. The original videos were public before the present evaluation; the fresh extension was generated separately. An anonymous release package, asset-

Original family	Videos
3d transformations	10
board games chess	10
causal physics intervention	10
containers object permanence	10
egocentric routes	10
home errands	10
home interior memory	5
home past state	30
industrial interior memory	5
multi person exchange	10
outdoor crossings	10
outdoor occluded delivery	20
physics memory	20
queues seating	10
street landmark history	35
tool use assembly	10

Table 11: Original collection by generator family, including development videos. The family-macro results in the main text use evaluation items only.

license inventory, and final dataset license are pending. We do not assert redistribution permission for assets without a completed license inventory.

From plan to scored item. We preserve the family-specific JSON descriptions rather than forcing all generators into one schema. A plan identifies entities and initial state, specifies the ordered events and camera motion, and records the question, option order, and answer logic. For the worked pressure example in Table 1, the query compares each gauge's reading at two bell cues and selects the greatest increase. The replacement-age example instead tracks the latest replacement of the plant in each fixed pot. Thus the queried state and temporal reference depend on the task family. The scoring procedure maps the resulting semantic answer to its labeled choice; changing option order requires the same mapping change during scoring. We retain all 500 full plans and their hashes in the supplied manifest. Separate frozen questions on the twelve extension histories support the diagnostics in Appendix H; they are additional questions on existing videos, not new scenes.

One level-4 development attempt failed because of service capacity before visual inspection. We applied the highest-level fallback and later completed the infrastructure retry without changing the selection.

Quality status. The extension's independent symbolic checks and Blender checks have completed. Two humans completed the qualitative solvability review in Appendix G; we report no quantitative human accuracy or inter-rater agreement. General minimal-evidence intervals are not available for the original collection. Camera variants should remain in the same partition as their source scene in future training or cross-validation uses. Results here describe a development/evaluation split for inference, not a training benchmark with independent generator families.

D Implementation and inference settings

Direct conditions. The six BF16 conditions use the following revisions: MiniCPM-V-4.6, `36f34a66`; GLM-4.6V-Flash, `411bb4d7`; InternVL3.5-8B-HF, `741a7d03`; Qwen3.5-9B, `c2022362`; Qwen3-VL-8B-Instruct, `0c351dd0`; and Qwen3-VL-8B-Thinking, `92f3c4b4`. The Qwen3-VL-32B-Instruct NF4 checkpoint is revision `48cfc649`, using NF4 language weights and BF16 vision weights. The supplementary analysis inventory retains the full revision strings. Uniform frame selection uses decoded timestamps spanning the video. The model's native processor can apply patch alignment or other resizing after the common preprocessing bound. Runs use a single RTX 5090 with 32 GB of memory. Latency depends on adapter and attention-kernel implementation and is not presented as a hardware-normalized model ranking.

The initial response may contain reasoning before its final JSON. Permitted JSON and presentation wrappers are parsed without accessing the key. A finalization request includes the same video observations, question, and preceding output; it instructs the model to return the required JSON structure. The constrained decoder restricts syntax, not the correct option. The final record preserves both the initial output and finalization attempts. A valid output can still be wrong.

Agent conditions. The inspection service decodes frames at requested times and logs the actual timestamps. Every displayed image is charged against B, including revisits and crops. The service also records pixels, elapsed time, calls, and submission. JPEG quality is 92. A fresh session prevents conversation history from the construction workflow being passed to the solver. The evaluator has access to the answer key only outside that session. This architecture addresses direct access during evaluation; it cannot prove a proprietary model's training provenance.

Quantized loader repair. An initial 32B development smoke test failed during model loading and produced no scored evaluation result. With Transformers 5.17.0 and bitsandbytes 0.50.2, the checkpoint's bare `visual` exclusion failed to protect the fully qualified vision module. We added `model.visual` to the loader exclusions and checked that all 448 quantized language layers had initialized quantization state and that vision layers remained BF16. Checkpoint bytes were unchanged. All reported 32B results come from completed reruns with this repair, identified by their worker hash and loader-adjustment record. This is a distinct precision condition, not a full-precision 32B result. The seven-model roster is a feasible selection of open-weight systems, not an exhaustive census.

E MATCHED-CONTROL STATISTICS

The frozen cohort contains four evaluation items per original family and no camera derivatives. The nine conditions pair three input controls with each of MiniCPM-V-4.6, InternVL3.5-8B, and Qwen3-VL-8B-Instruct. Each uniform reference is extracted from that model's full original run. We verify exact item matching and that each shuffled condition uses the same timestamp multiset as its uniform reference. Shuffled images are relabelled with chronological presentation timestamps. The question-only and middle-frame conditions supply zero and one image, respectively; their native processor paths and input costs differ from 64-frame video.

Model	Control	Δ (pp)	95% paired CI	Gain/loss	p_{Holm}
MiniCPM-V-4.6	question only	+1.6	[-7.8, 10.9]	5/4	1.000
MiniCPM-V-4.6	middle	+0.0	[-7.8, 7.8]	3/3	1.000
MiniCPM-V-4.6	shuffle	+4.7	[-3.1, 12.5]	5/2	1.000
InternVL3.5-8B	question only	+3.1	[-9.4, 15.6]	9/7	1.000
InternVL3.5-8B	middle	+4.7	[-7.8, 17.2]	10/7	1.000
InternVL3.5-8B	shuffle	+4.7	[-4.7, 15.6]	7/4	1.000
Qwen3-VL-8B-I	question only	+7.8	[-3.1, 18.8]	9/4	1.000
Qwen3-VL-8B-I	middle	+12.5	[0.0, 23.4]	12/4	0.691
Qwen3-VL-8B-I	shuffle	+1.6	[-9.4, 10.9]	6/5	1.000

Table 12: Control minus matched uniform accuracy on 64 items, in percentage points. Gain/loss counts items corrected/broken by the control. Bootstrap intervals are unadjusted percentile intervals; exact McNemar tests use Holm correction across all nine comparisons. A percentile endpoint of zero does not imply a significant corrected test.

Table 12 and Figure 6 show the same paired original-control comparisons numerically and graphically.

F STATISTICAL ANALYSIS DETAILS

The CPU script reads the final per-item record in each selected condition and scores it against the controller's key. It rejects incomplete headline conditions. No questions, answer keys, source scenes, or raw private traces are copied into the aggregate paper snapshot. The snapshot includes input content hashes and a UTC freeze time. Rebuilding tables from this snapshot does not read a live evaluator output.

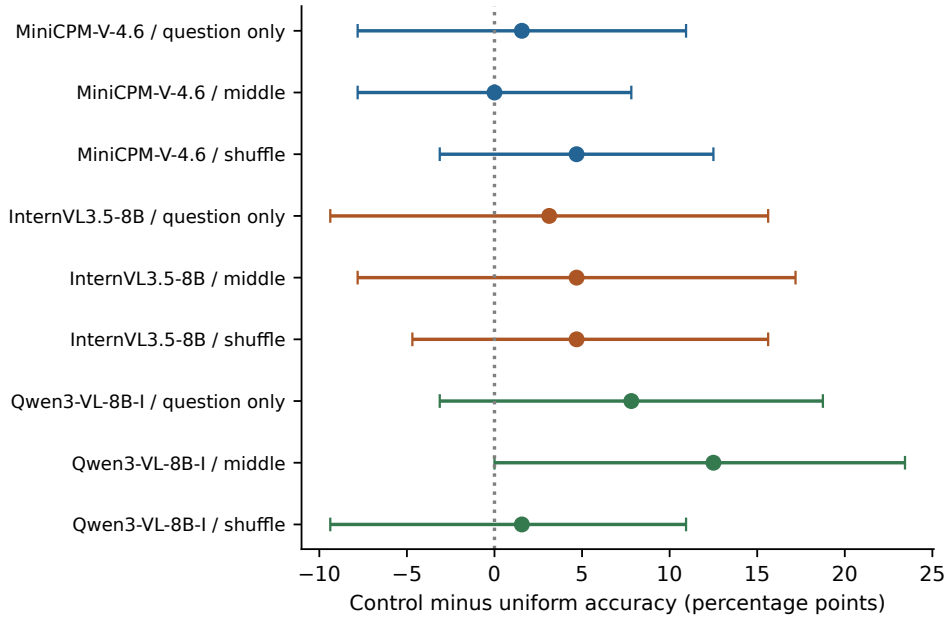


Figure 6: Paired differences for completed controls. Positive values favor the control over chronological uniform video. Horizontal lines are 95% item-bootstrap intervals; they do not establish equivalence when overlapping zero. All Holm-adjusted exact-test p-values are at least 0.691.

Wilson intervals use $z = 1.959963984540054$. Paired bootstrap intervals resample item differences with replacement 10,000 times using seed 260916. Exact McNemar tests condition on discordant pairs. Holm adjustment is applied separately to the three original agent comparisons, the 21 completed direct-model comparisons, and the nine matched-control comparisons. Item dependence remains after dropping camera derivatives. The 195 items contain 180 normalized semantic-history groups; the 190-item camera-excluded subset still contains 180. Two evaluation items share exact histories with development. Excluding these gives Astra scores of 152/193, 187/193, and 188/193 at the three budgets. An additional paired bootstrap resampling all members of exact history groups (10,000 draws, seed 20260919) gives 95% intervals of [12.2,24.1], [12.8,24.5], and [−1.0,2.5] points for the 16-to-64, 16-to-256, and 64-to-256 changes. Original item-based intervals and exact tests use a working independence assumption. History grouping does not eliminate shared-template dependence. Full exclusion sensitivities are supplied in `analysis/history_sensitivity.json`; the separate adaptation split is history-disjoint.

Confidence diagnostics in the snapshot score the predicted probability that the chosen answer is correct using a binary Brier score. A four-class Brier score would require probabilities for all options, which the response contract does not provide. We do not fit a mixed-effects difficulty model to the small adaptive extension: levels change correlated factors, and perfect development accuracy would produce unstable estimates.

Placement-window proxy. For each extension relocation e, let $W_e = [t_{\text{carry end}}, t_{\text{event end}}]$. This includes the placement animation and 0.7 seconds of settling. Given the set of displayed timestamps T_i, the item's proxy is

$$C_i = \frac{1}{|\mathcal{E}_i|} \sum_{e \in \mathcal{E}_i} \mathbf{1}\{T_i \cap W_e \neq \varnothing\}. \tag{2}$$

We average C_i over the 12 validation items. Uniform coverage uses B equally spaced timestamps from zero to one decoded frame before the video ends. This proxy does not account for occlusion, image resolution, alternative sufficient evidence, or whether a particular relocation contributes to the answer. It must not be reported as decisive-evidence recall.

G HUMAN REVIEW AND FURTHER VALIDATION

Completed offline review. Two human reviewers evaluated a random sample of 50 generated videos and their associated puzzles, representing 10% of the 500-video dataset. They conducted the review offline and judged all 50 video–puzzle pairs sufficiently clear and informative for a human to solve the corresponding puzzle. This qualitative check supports the usability of the reviewed sample. We do not interpret the verdict as a measured human accuracy score or as validation of the remaining 450 items.

Proposed quantitative follow-up. Recruit independent participants who have not seen generator code or answers. Each participant should answer the same video/question items without feedback, record confidence, and flag unclear moments with timestamps. The planned primary condition permits one viewing and at most one optional replay; an unrestricted-replay condition should use a separate assignment. Randomize item order and balance scene families across sessions to limit fatigue.

The existing local review interface collects selected answers, confidence, ambiguity notes, elapsed time, and play events. Play events are not a reliable count of complete replays because pausing and resuming also produces them. Accurate exposure analysis requires seek and playback-interval instrumentation. Consent, compensation, and institutional review status must be documented before recruitment.

Report per-item accuracy, participant-level uncertainty, agreement, and the number of exclusions. Have at least three independent ratings per retained item where feasible. Ambiguity decisions should be made without seeing model correctness; preserve the full collection score and an independently adjudicated subset score. Do not choose examples for the human study because a particular model failed on them, or change a validation answer after inspecting model predictions without versioning the item.

H DIAGNOSTIC FOLLOW-UPS AND REMAINING EXPERIMENTS

The follow-ups below adapt reliable intermediate representations from MentisOculi to observed histories (Zeller et al., 2026). They are exploratory input interventions, not internal mechanistic evidence. Strict causal scene pairs, activation interventions, and quantitative human accuracy evaluation remain unrun.

H.1 COMPLETED FOLLOW-UP COHORT AND RESULTS

These 388 completed records are separate from the headline snapshot: 48 oracle-input answers, 60 native timing/record answers, 108 native prefix/control records, 108 Astra answers, and 64 generation-budget answers. The first four groups use 12 existing validation videos with new questions, selected from four symbolically verified candidates per history using a fixed seed before inference. Reference ranks span 1–3 and retrospective depths 0–2. They are not the original challenge questions and do not establish increasing difficulty. Correct positions are A=1, B=1, C=5, D=5; a constant majority position scores 5/12. All conditions retain the same questions, except the explicit option permutations, which are scored after semantic remapping.

Table 5 contains only conditions completed for both systems. Table 13 reports the additional completed diagnostics under their respective interfaces. We did not run Astra counterparts for native uniform-video, event-aligned, shuffled-frame, or three option-rotation conditions. We also did not run native counterparts for the two agent replay conditions. In particular, selected replay requires an observation-selection interface absent from the native protocol. These eight unrun comparisons remain unmeasured; no missing result is imputed or scored as zero.

Inputs and isolation. Qwen3-VL-8B-Instruct uses 64 frames with input area capped at 448^2, a 768-token reasoning allowance, and constrained finalization. Uniform and event-aligned inputs have the same frame count. Transcripts contain all events; state inputs contain complete state histories. Both are privileged symbolic inputs. In native timing conditions, observation occurs with or without

Condition	Evaluated system	Correct/total
Uniform video	Qwen3-VL-8B-I	2/12
Event-aligned frames	Qwen3-VL-8B-I	3/12
Shuffled frames	Qwen3-VL-8B-I	2/12
Option rotation 1	Qwen3-VL-8B-I	2/12
Option rotation 2	Qwen3-VL-8B-I	4/12
Option rotation 3	Qwen3-VL-8B-I	3/12
Selected replay (up to 32 extra frames)	Astra	12/12
32 additional uniform frames	Astra	12/12

Table 13: Additional completed follow-up diagnostics, correct/total. Each row names the system actually evaluated. Cross-interface counterparts were not run; the native model is not equipped with the agent's selected-replay tools. We retain these results separately rather than implying a matched comparison.

the question, followed by a 512-token record and an answer with visual context retained. Bounded-record answering instead receives only the record and question in fresh messages, with no images.

Astra receives 64 chronological observations in eight calls. Question-after conditions reveal the question only after those observations. Replay permits up to 32 selected additional frames; the matched allowance condition supplies 32 fixed additional uniform frames. Actual selected replay uses 77–92 total frames, compared with 96 in the fixed condition; these are matched allowances, not matched realized costs. The call cap is 16. For bounded records, a question-blind observer sees 64 frames; a fresh answerer receives only the retained text, question, and choices. The 128/512-token caps use a Qwen tokenizer proxy and therefore are operational text budgets, not Astra-native token measurements. Trace audits confirm no image calls by these fresh answerers and enforce the caps. The no-record condition supplies only the question and choices.

Interpretation and state probes. Astra answers all 12 questions correctly in both timing conditions, both bounded-record conditions, replay, additional-uniform, transcript, and state conditions, compared with 1/12 without a record. This shows useful information in the external records on this small cohort, while leaving replay benefits unresolved at ceiling. Native oracle-input scores remain low (Table 5). Four fresh prefix probes per history yield 51/144 correct object locations and 1/48 completely correct three-object states. These are output-state checks; they do not reveal latent state representations. The small, answer-imbalanced cohort and changed input modalities limit causal interpretation.

Prefix-state protocol. We choose four checkpoints at approximately 25%, 50%, 75%, and 100% of each history's recorded event boundaries, rather than equal fractions of video duration. We sample 64 frames uniformly from the start to each checkpoint and ask a fresh session for each object's current station, without supplying the final puzzle. The prompt also requests activity counts; the reported state metrics score only the three object locations. We compare each location with the symbolic checkpoint state and count a complete state as correct only when all three locations match. Twelve histories yield 48 state reports and 144 location predictions. These separate sessions do not test retention across successive probe calls.

Generation-budget check. On the same 32 original evaluation items, Qwen3-VL-8B-Thinking scores 10/32 at a 2,048-token initial allowance and 11/32 at 8,192. Initial truncations fall from 13 to five; mean initial output length rises from 1,585 to 2,767 tokens. Both conditions retain constrained finalization. Reducing truncation therefore produces only a one-answer observed gain here; the experiment is too small to establish equivalence or an optimal budget. Frozen aggregates and protocol checks are supplied in `analysis/followup_results.json`.

Figure 4 places the answer-level interventions beside state-output accuracy. The low native prefix scores motivate supervision of intermediate state, but they do not show whether a visual, binding, or retrieval error caused a particular answer. Astra's symbolic and record conditions reach ceiling on these twelve histories, leaving the relative benefits of those representations unresolved.

H.2 Experiments supported by the scene plans

The completed interventions compare observed video, symbolic descriptions, and explicit external records. Following MentisOculi's input and intermediate-representation comparisons, we treat differences as effects of supplied information and representation, not additive estimates of perception and reasoning error (Zeller et al., 2026). The scene plans permit a more targeted follow-up: score a model-produced ledger, replace one prespecified incorrect entry, and compare the answer with an equally sized irrelevant-entry edit. Correct-ledger/wrong-answer and wrong-ledger/correct-answer outcomes would help characterize the interface between explicit state output and answering. We have not run that rescue experiment.

To test interference, new matched renders should vary retention delay and intervening distractor updates independently, holding decisive-event visibility and answer distributions fixed. A separate minimal pair could change one decisive event while preserving unrelated actions. Camera-only and irrelevant-event edits would provide answer-invariant controls. The existing camera derivatives and authored difficulty levels do not supply these causal comparisons. Likewise, the current joint frame/call budgets cannot separate the contribution of frame allowance from tool-call allowance.

Activation probes and patching require additional controls before supporting claims about internal memory. A suitable study would split by full history, use low-capacity probes with shuffled-label baselines, select layers on development data, and compare target-state patches with irrelevant-state and norm-matched controls. None of these activation experiments has been run. The accompanying historical analysis agenda records these designs; it is not evidence of completed measurements.

I Fine-tuning setup and controls

Partition and scope. This exploratory adaptation experiment uses a separate retrospective partition of the already inspected collection. The first 200 original videos supply 120 training, 24 development, and 56 test videos, containing 112, 23, and 53 exact semantic-history groups. Grouping ignores presentation text, timestamps, camera paths, and the selected question while retaining initial states, ordered state-affecting events, family, and physical parameters. Exact history and video hashes cannot cross active partitions; five camera derivatives are excluded. This check does not rule out every semantic isomorphism or shared-template dependency. The ten roofed-home/workshop scenes and all 24 later history-binding scenes form a 34-video transfer set across four family labels absent from training. This is transfer within the renderer. No external real-video transfer evaluation was run.

Some adaptation training videos were previously inference-evaluation examples. The pilot therefore has its own split and matched baseline; it does not replace the original benchmark scores. Test and transfer histories were held out from adapter updates and checkpoint selection, although they had already been inspected during benchmark development. Alternative questions and visual derivatives inherit their parent history's partition. Test input manifests omit answers, and the scoring key remains outside model inputs. No answer-bearing paper gallery or symbolic state transcript is supplied during training.

Model, input, and optimization. The backbone is Qwen3-VL-2B-Instruct, revision `89644892`, with BF16 backbone computation. We train 6,422,528 LoRA parameters on the language attention query, key, value, and output projections, with rank 16, scaling parameter $\alpha = 32$, and dropout 0.05. The vision encoder, visual projector, and remaining backbone weights are frozen. Each video contributes 32 uniformly sampled frames with their timestamps, aspect ratio preserved, and input area capped at 336^2 pixels per frame. Native processing disables additional implicit frame sampling. The context cap is 12,288 tokens; inputs are rejected rather than silently truncated.

Each of the 120 training videos contributes one question and its four options. Supervised loss is masked on all prompt and visual tokens and applies only to the correct answer letter and termination token. There are no rationale or intermediate-state targets. We use AdamW with learning rate 10^{-4}, weight decay 0.01, microbatch one, accumulation over eight examples, gradient clipping at norm one, and a cosine schedule with three warmup updates. Three epochs give 360 example presentations and 45 optimizer updates per arm. Gradient checkpointing is enabled. The question-only arm has

the same examples, labels, update count, and seed but omits frames; its input-token and compute costs are much lower.

Selection and validation. A discarded two-example development smoke run checks label masking, nonzero adapter gradients, and loss reduction before the pilot. Each epoch is evaluated on all 24 development examples using the arm's training input modality. The checkpoint with highest development accuracy is selected, with ties going to the earliest epoch: epoch two for video-answer training (8/24 with video) and epoch one for question-only training (9/24 without video). Adapter hashes are frozen before held-out inference. All arms use the identical single-letter A–D output constraint. Only seed 17 completed the full comparison; seeds 29 and 43 were deferred under the time budget. Results therefore provide no estimate of training-seed variability.

Both three-epoch jobs ran on one RTX 5090 with 32 GB of memory, using PyTorch 2.9.1, Transformers 5.17.0, and PEFT 0.18.1. Recorded training-loop times, including epoch-level development evaluation but excluding model loading and held-out evaluation, are 2.49 hours for video supervision and 49.1 seconds for question-only supervision. Peak allocated GPU memory is approximately 7.99 GB and 4.61 GB, respectively. These costs are specific to this implementation.

	Test ($n = 56$)			Family transfer ($n = 34$)		
Arm	Video	No video	Shuffled	Video	No video	Shuffled
Untouched base	14/56	16/56	13/56	11/34	10/34	9/34
Video-answer SFT	12/56	12/56	12/56	8/34	8/34	8/34
Question-only SFT	11/56	11/56	–	8/34	8/34	–

Table 14: Held-out pilot input controls, correct/total. The shuffled condition permutes the same 32 frames and retains chronological presentation timestamps. The two dashes mark shuffled-frame evaluations of the question-only SFT adapter that we did not run on either split; we evaluated that adapter only with chronological video and no video. These cells are unmeasured, not zero scores or failed runs. Question-only training evaluated with video is a modality-transfer control.

Interpretation. The video-trained model has the same aggregate held-out scores with video, question-only input, and shuffled frames (Table 14); this does not imply identical per-item predictions. A subsequent diagnostic using the frozen adapters compares all development examples and a fixed hash-selected subset of 24 training videos. On that training subset, full-video accuracy rises from 2/24 for the base to 10/24 for the video-trained adapter, which also scores 10/24 without video. These checks support further investigation of answer-distribution adaptation and weak visual supervision; they do not identify an internal failure mechanism.

The video-minus-base item-accuracy change is -3.6 points on the 56 test videos and -8.8 points on the 34 transfer videos. Corresponding 95% intervals are $[-20.0, 14.0]$ and $[-32.4, 14.7]$ points, using 10,000 bootstrap resamples of exact history groups, retaining all group members (seed 20260919). The intervals describe history-sampling uncertainty, not training-seed uncertainty. Supervision of object identities, intermediate state updates, and event order may better exploit the scene programs, but improved held-out performance from such supervision has not been measured.

Post-hoc answer-position diagnostic. Training labels have counts A=21, B=47, C=25, D=27. Test labels have counts (12,11,16,17), and transfer labels (11,8,8,7). Predicting the training-majority letter B always would score 11/56 and 8/34. With video, the video-trained adapter predicts B on 55/56 test and all 34 transfer items; the base predicts A on 80/90 combined items, and the question-trained adapter predicts B on 77/90. Thus answer-position adaptation is a plausible explanation for the weak visual-control differences, although these counts do not establish an internal mechanism. On a fixed eight-item test subset evaluated under all four cyclic option orders, semantic-answer consistency across every order is 2/8 for the base, 0/8 for video training, and 3/8 for question-only training. This is a post-hoc diagnostic, not a checkpoint-selection criterion.

I.1 Balanced-label follow-up

After observing the original adapter's answer-position collapse, we froze a follow-up protocol before its runs. We kept the same 120 training videos, 24 development videos, 56 test videos, and 34 family-

transfer videos. We permuted each training item's choices so that each epoch has exactly 30 correct A, B, C, and D labels. During epochs 1, 2, and 3, every item receives an additional cyclic option shift of 0, 1, and 2 positions; the supervised letter changes with the permutation. An independent check confirms semantic-answer preservation and byte-identical development/test inputs. No new scene is added, and the test histories were already used in the original pilot.

We train separate video and question-only adapters with the original backbone revision, rank-16 LoRA, three epochs, learning rate 10^{-4}, effective batch size eight, and seed 17. Development accuracy in the training modality selects the epoch, with earliest ties. The unadapted baseline reuses frozen predictions on identical inputs. Held-out evaluation uses the original option order plus the same fixed eight-item cyclic-order diagnostic. Balanced training changes option placement and its diversity together, so it is not an isolated test of either intervention.

Arm	Input	Test	Transfer	Test option consistency
Base	video	14/56	11/34	2/8
Base	question only	16/56	10/34	—
Base	shuffle	13/56	9/34	—
Balanced question SFT	video	12/56	9/34	2/8
Balanced question SFT	question only	13/56	6/34	—
Balanced video SFT	video	18/56	9/34	1/8
Balanced video SFT	question only	17/56	10/34	—
Balanced video SFT	shuffle	14/56	9/34	—

Table 15: Balanced-label follow-up, correct/total. Test and transfer remain excluded from gradient updates and checkpoint selection but have been evaluated before this experiment. Option consistency is semantic agreement across all four cyclic orders on eight fixed test items; it does not require correctness. We ran this option-order diagnostic only with chronological video. Dashes in the consistency column therefore mean unmeasured for question-only and shuffled inputs, not zero consistency or failed evaluations.

Table 15 reports all completed input conditions. The base and video-trained models receive video, question-only, and shuffled inputs; the question-trained model receives video and question-only inputs. History-cluster intervals use 10,000 resamples with seed 20260924; they do not quantify training-seed variation. Intermediate-state supervision, a larger history-disjoint training set, and a fresh validation cohort would provide stronger tests of training utility.

Answer-position bias persists: the balanced video adapter chooses C on 75/90 test and transfer questions in their original option order. Its semantic answer is consistent across all four cyclic orders on only 1/8 diagnostic items. Balanced training labels alone did not remove this sensitivity.

J Later-cohort construction and analyses

We added three cohorts after the initial 239 videos. Diverse-100 covers stack adjacency, accumulated watering quantities, repainting, lamp states, rover headings, concealed cargo, shunting order, weight differences, cyclic book order, and branching channels. Mechanism-100 covers conserved transfers, coupled gears, recipe-set intersections, AND/XOR relays, magnetic pole pairing, elevator capacity, woven-band alternation, percussion counts, route displacement, and numbered-tag provenance. Each contains ten videos per family. Final-61 adds continuous activity duration (11 videos) and five ten-video families: pressure changes between checkpoints, priority scheduling with waiting-time ties, replacement age, directed event transitions, and ideal pulley stages.

The three cohorts last 46–74, 67–95, and 74–127 seconds. We render at 1280×720, 24 fps, with Eevee 64 samples and H.264 CRF 17. These scripted exhibits follow explicit symbolic rules; pulley ratios and gear updates are not validated force simulations. Within each cohort, histories have unique recorded signatures and agree with separately reconstructed transitions. This does not rule out semantic equivalence across generators. The current audit rechecks each answer program and the hash of every encoded video against its completion record.

We authored questions before their model runs and included every generated item. The two 100-video cohorts balance answers at 25 per position; Final-61 uses (16,15,15,15). Each requested

historical answer differs from the corresponding final-state answer under the generator's query convention. Difficulty labels are design estimates. Earlier outcomes informed later generator choices, so these cohorts are sequential exploratory extensions. No later-cohort outcomes select fine-tuning checkpoints. The small adaptation studies use their own original-history split.

Matched controls and uncertainty. We select four items per family by an answer-blind hash: 40, 40, and 24 items. Correct-position counts A/B/C/D in these subsets are (12,13,8,7), (10,10,8,12), and (6,5,9,4). Thus a majority-position predictor scores 13/40, 12/40, or 9/24. Table 16 reports the chronological reference on exactly the same items as each control, together with control-minus-video differences. We retain all nine comparisons per cohort, including negative and inconclusive differences. Holm adjustment covers nine tests within each cohort; it does not cover every exploratory comparison in the manuscript.

Table 16: Later-cohort matched controls. Delta is control minus chronological video in percentage points; intervals are unadjusted paired bootstrap 95% intervals. Every cell reports its denominator.

Cohort / model	Control	Video	Control	Δ (pp)	95% interval	p_{Holm}
Diverse / MiniCPM-V-4.6	Text	14/40	12/40	-5.0	[-15.0,5.0]	1
Diverse / MiniCPM-V-4.6	Last	14/40	13/40	-2.5	[-12.5,7.5]	1
Diverse / MiniCPM-V-4.6	Shuffle	14/40	13/40	-2.5	[-15.0,10.0]	1
Diverse / Qwen3.5-9B	Text	21/40	8/40	-32.5	[-50.0,-15.0]	0.0398
Diverse / Qwen3.5-9B	Last	21/40	9/40	-30.0	[-50.0,-10.0]	0.0945
Diverse / Qwen3.5-9B	Shuffle	21/40	18/40	-7.5	[-25.0,12.5]	1
Diverse / Qwen3-VL-8B-I	Text	16/40	14/40	-5.0	[-25.0,15.0]	1
Diverse / Qwen3-VL-8B-I	Last	16/40	8/40	-20.0	[-37.5,-2.5]	0.538
Diverse / Qwen3-VL-8B-I	Shuffle	16/40	13/40	-7.5	[-20.0,5.0]	1
Mechanism / MiniCPM-V-4.6	Text	10/40	11/40	+2.5	[-5.0,10.0]	1
Mechanism / MiniCPM-V-4.6	Last	10/40	9/40	-2.5	[-12.5,7.5]	1
Mechanism / MiniCPM-V-4.6	Shuffle	10/40	9/40	-2.5	[-12.5,5.0]	1
Mechanism / Qwen3.5-9B	Text	21/40	15/40	-15.0	[-37.5,7.5]	1
Mechanism / Qwen3.5-9B	Last	21/40	2/40	-47.5	[-62.5,-32.5]	3.43e-05
Mechanism / Qwen3.5-9B	Shuffle	21/40	13/40	-20.0	[-35.0,-5.0]	0.309
Mechanism / Qwen3-VL-8B-I	Text	12/40	8/40	-10.0	[-25.0,5.0]	1
Mechanism / Qwen3-VL-8B-I	Last	12/40	9/40	-7.5	[-20.0,5.0]	1
Mechanism / Qwen3-VL-8B-I	Shuffle	12/40	6/40	-15.0	[-27.5,-2.5]	0.492
Final / MiniCPM-V-4.6	Text	5/24	6/24	+4.2	[0.0,12.5]	1
Final / MiniCPM-V-4.6	Last	5/24	6/24	+4.2	[-8.3,16.7]	1
Final / MiniCPM-V-4.6	Shuffle	5/24	8/24	+12.5	[0.0,25.0]	1
Final / Qwen3.5-9B	Text	4/24	11/24	+29.2	[4.2,54.2]	0.589
Final / Qwen3.5-9B	Last	4/24	3/24	-4.2	[-20.8,12.5]	1
Final / Qwen3.5-9B	Shuffle	4/24	6/24	+8.3	[-12.5,29.2]	1
Final / Qwen3-VL-8B-I	Text	8/24	11/24	+12.5	[-12.5,37.5]	1
Final / Qwen3-VL-8B-I	Last	8/24	9/24	+4.2	[-20.8,29.2]	1
Final / Qwen3-VL-8B-I	Shuffle	8/24	10/24	+8.3	[-12.6,29.2]	1

We recompute paired intervals from 10,000 item/history resamples with seed 20260924 and use exact McNemar tests. The portable aggregate `analysis/later_audit.json` retains all scores, paired tests, source hashes, label counts, and trace summaries. The earlier paired snapshots retain their original seed. Distinct exact histories still share generator structure; the intervals use a working item-independence assumption. Cross-cohort differences change task mechanisms and visuals together and cannot establish a causal difficulty effect.

Error concentration and shared failures. Figure 7 shows why one aggregate score can hide different weaknesses. For example, GLM scores 9/10 on repainting in Diverse-100 but 2/10 on cyclic book order and branching channels. On Mechanism-100, Qwen3.5 scores 8/10 on both recipe overlap and package provenance, compared with 2/10 on coupled gears. On Final-61, the relative advantage shifts again: GLM scores 8/10 on priority dispatch while InternVL scores 2/10, but InternVL scores 6/10 on checkpoint trends while GLM scores 1/10. These are descriptive comparisons over the full family inventory, not selected causal tests of particular skills.

Figure 8 compares all five design levels within cohort. We retain the complete curves rather than selecting a threshold from observed errors. Event counts, reference placement, and visual settings co-vary, so a downward segment cannot be interpreted as a forgetting rate. Independent variations of delay, intervening updates, and evidence visibility would be needed to distinguish those explanations.

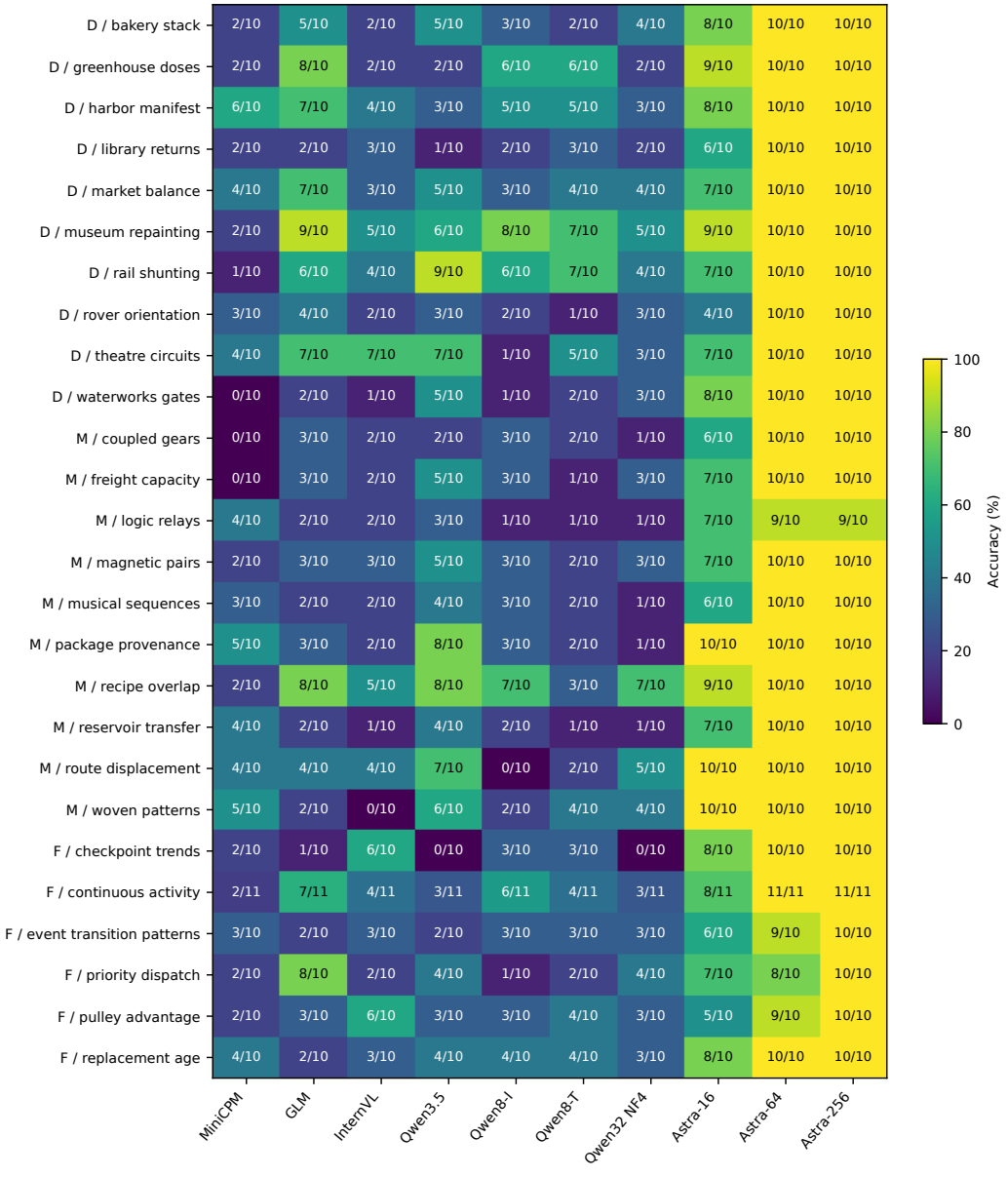


Figure 7: All 26 later-cohort families and ten primary conditions. D/M/F denote Diverse-100, Mechanism-100, and Final-61. Cells show correct/total, with color indicating accuracy. Rows have ten items except continuous activity, which has eleven. We show every family, including those where open models perform relatively well. Shared visual and event structures and small cell sizes limit model ranking and causal interpretation.

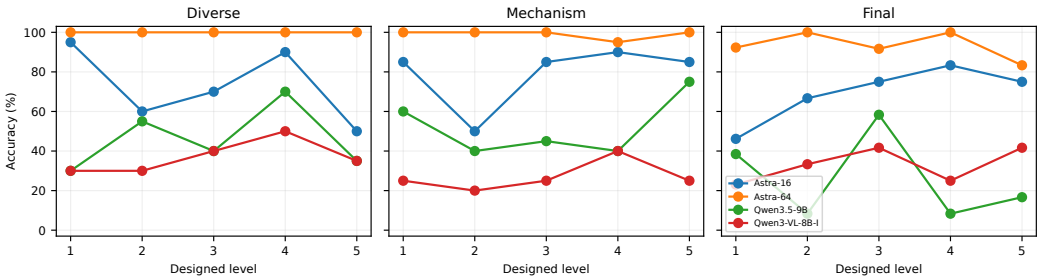


Figure 8: Accuracy by authored level for four fixed conditions. Each point contains 20 items in each 100-video cohort; Final-61 contains 13 at level 1 and 12 at each other level. Levels jointly vary event histories and rendering parameters. The lines summarize descriptive associations and do not validate a monotonic difficulty scale.

Generation and inspection costs. On Final-61, Qwen3.5 reaches the initial token cap on 49/61 items and uses finalization on 55/61. Qwen3-VL-8B-Thinking reaches the cap and uses finalization on 55/61. GLM does so on 21/61 and 23/61, respectively. The 32B condition has no initial truncations but scores only 16/61. These observations show that truncation alone does not account for all low scores; they do not establish how an uncapped model would perform. MiniCPM uses finalization on all 61 responses, so its score describes that complete response protocol. Agent actual-use curves appear in Figure 2; allowance is distinct from realized visual cost.

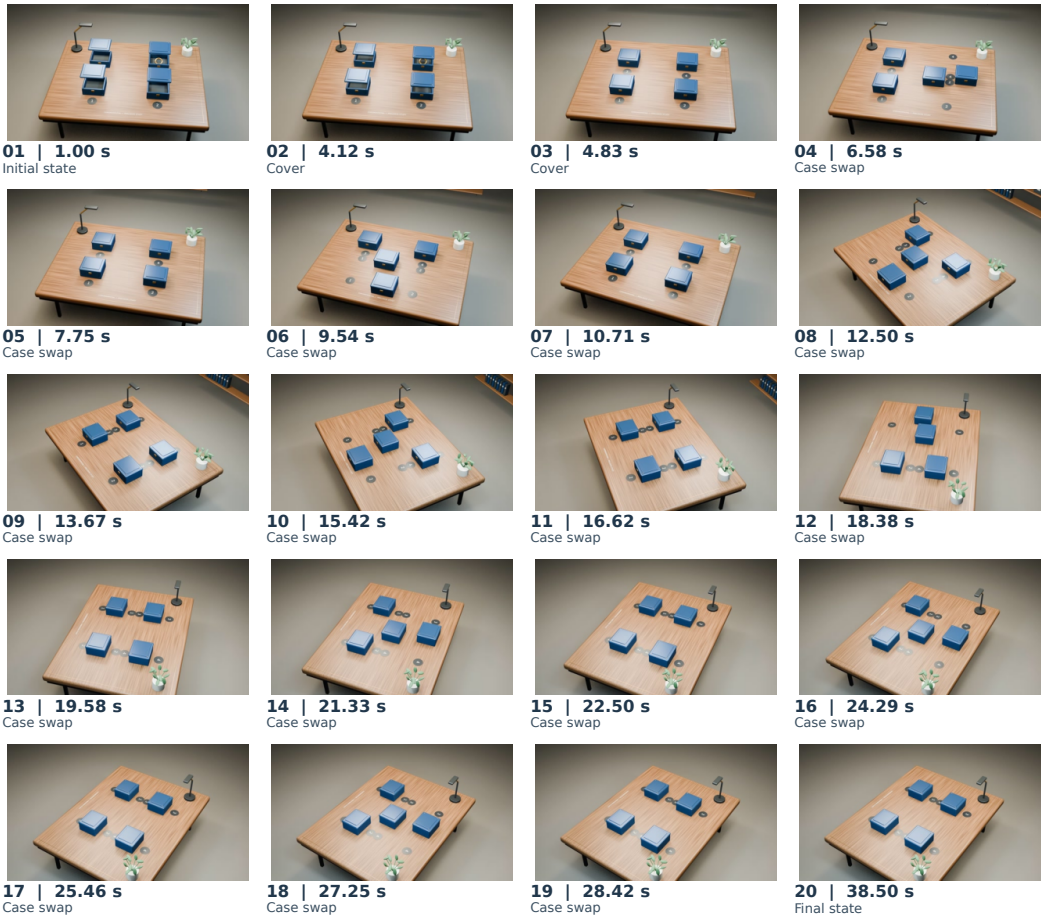
K VIDEO EXAMPLES WITH ANSWER KEYS

The following 20 pages illustrate all 16 original families and four extension development scenes. Each page shows 20 chronological frames in a four-column, five-row grid. Timestamps identify decoded frames; action labels come from the scene scripts. Selection uses event boundaries and query references, without model predictions. These summaries may omit intermediate actions and do not establish human solvability. Six original evaluation items are illustrated after the result freeze; extension validation items are excluded. These answer-bearing pages are documentation, not solver inputs. The full videos remain the evaluation stimuli.

Example	Scene focus	Page
01	Hidden contents and case identity	30
02	Rook tracking on a rotating board	31
03	Turns and relative direction	32
04	Keyring transfers between drawers	33
05	Object ownership across exchanges	34
06	Crossing paths and hidden motion	35
07	Remembering assembly order	36
08	Tracking a face through 3D rotations	37
09	Identity through queue rearrangements	38
10	Remembering a mechanical configuration	39
11	A past location at a light cue	40
12	Landmark history with moving viewpoints	41
13	Deliveries behind an occluder	42
14	Remembering a physical calibration	43
15	Activities inside a roofed home	44
16	Item history inside a workshop	45
17	Binding repeated activities to object history	46
18	Binding workshop deliveries to earlier states	47
19	Binding repeated activities to object history	48
20	Binding workshop deliveries to earlier states	49

K.1 Example 01: Hidden contents and case identity

Original scene 010; Original development; 39 s; assigned level 4 (not human-calibrated).



Puzzle. At the end, which numbered station holds the case containing the brass ring?

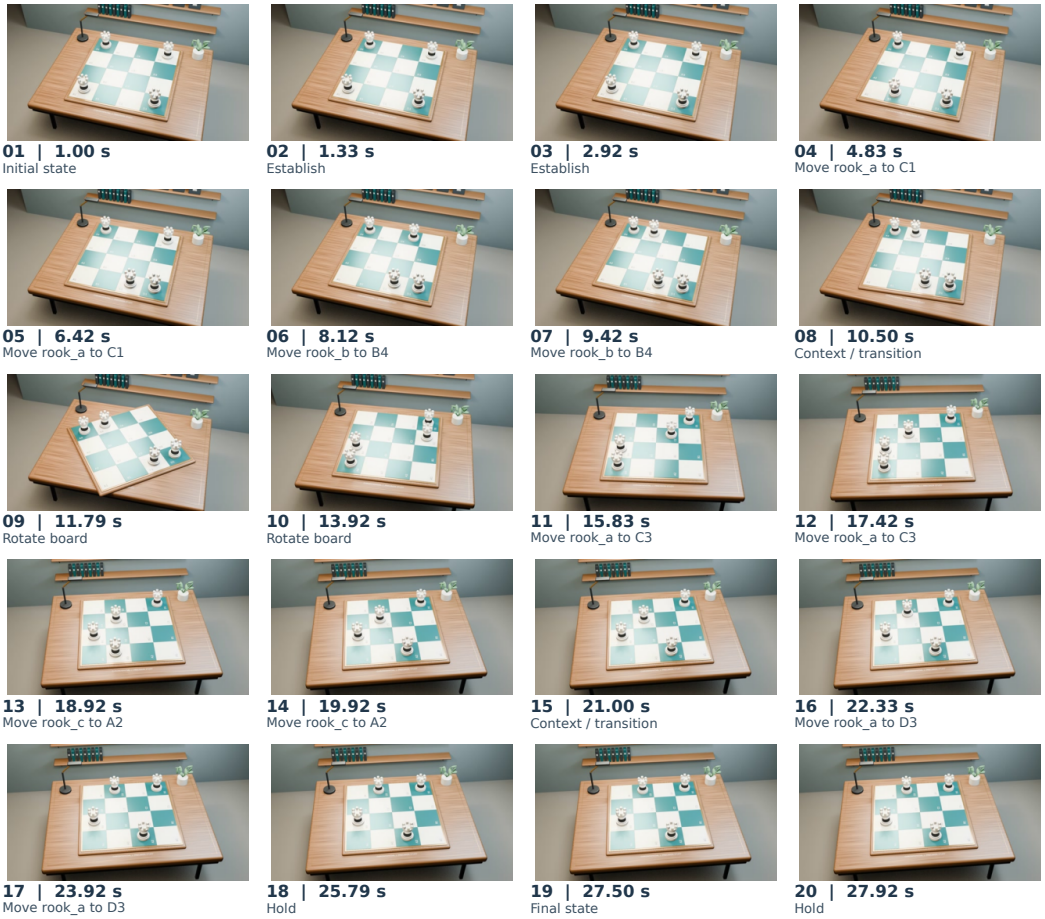
Track the ring before the matching cases close. Station numbers stay on the table; contents stay inside their cases.

A. Position 3
B. Position 4
C. Position 2
D. Position 1

Correct answer: B (Position 4). Replaying the 9 scripted events with persistent identities gives Position 4. The images show action boundaries and intervening motion; the stored answer is independently checked by the renderer-independent event interpreter.

K.2 Example 02: Rook tracking on a rotating board

Original scene 002; Original development; 28 s; assigned level 3 (not human-calibrated).



Puzzle. Which square holds the rook that started on A1 at the end of the video?

Track the rook that starts on A1. The coordinates rotate with the board. All four rooks look alike; no chess knowledge is needed.

A. D3 B. B4
C. B3 D. A2

Correct answer: A (D3). The rook follows A1 → C1 → C3 → D3. The board turns with its coordinate labels, so the final square is D3.

K.3 Example 03: Turns and relative direction

Original scene 041; Original development; 59 s; assigned level 5 (not human-calibrated).



Puzzle. Relative to the final forward direction, which way were you facing at the start?

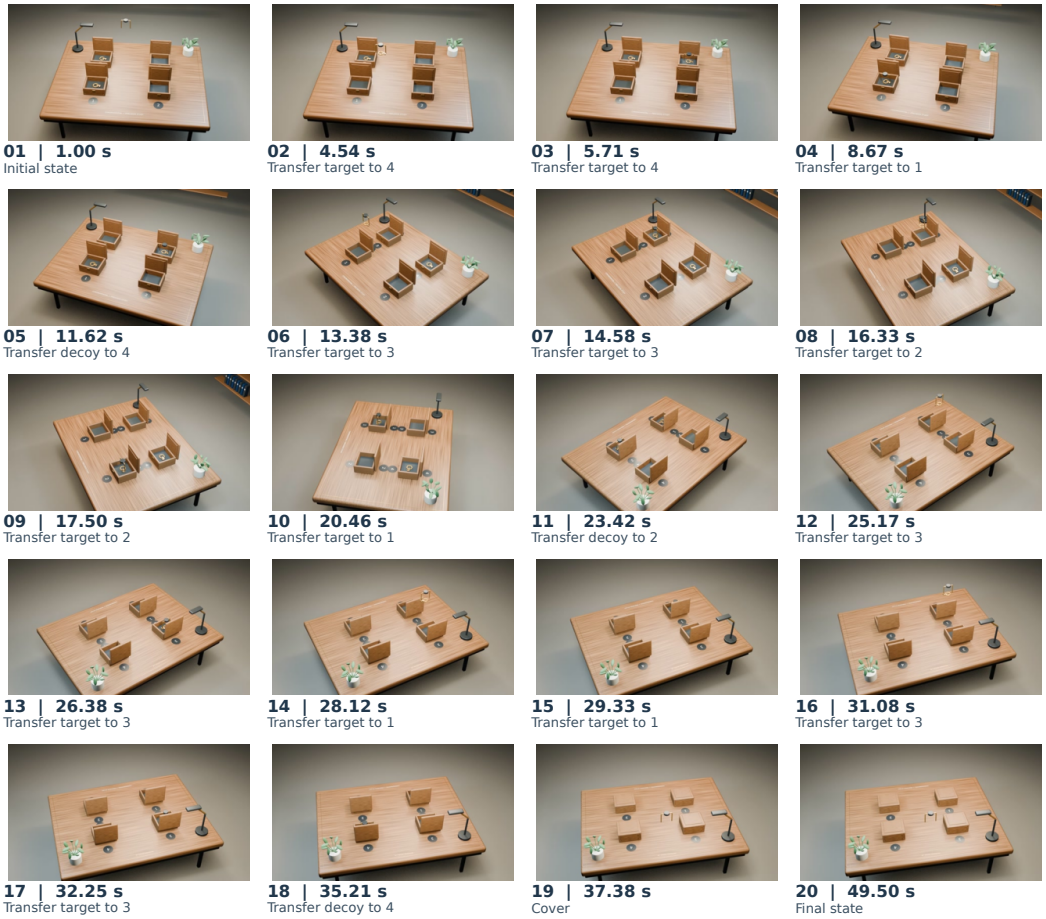
You are viewing the route through the traveller's eyes. Remember the initial view and track each turn; backward movement does not turn the traveller.

A. To the right
B. Straight ahead
C. To the left
D. Directly behind

Correct answer: A (To the right). Turn right through 90 degrees. Turn right through 90 degrees. Turn right through 90 degrees. Turn right through 90 degrees. Turn right through 90 degrees. Turn left through 90 degrees. Turn left through 90 degrees. Turn left through 90 degrees. Turn left through 90 degrees. Turn left through 90 degrees. Turn left through 90 degrees. Therefore the answer is To the right.

K.4 Example 04: Keyring transfers between drawers

Original scene 023; Original development; 50 s; assigned level 5 (not human-calibrated).



Puzzle. Which drawer now contains the keyring that began in drawer 1?

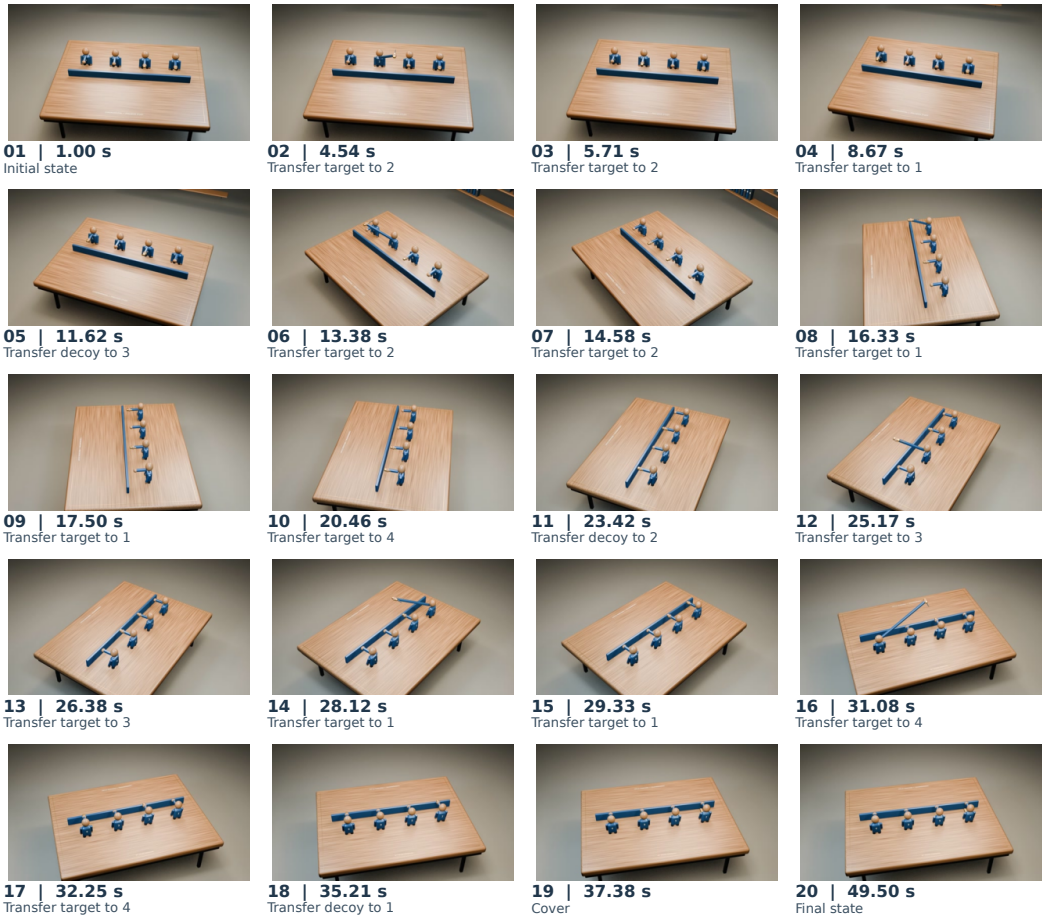
Track the keyring initially in drawer 1. The matching spare is a distractor. Drawer numbers stay fixed.

A. Drawer 1
B. Drawer 4
C. Drawer 3
D. Drawer 2

Correct answer: C (Drawer 3). The original keyring is visibly moved from drawer 1 to 4. The original keyring is visibly moved from drawer 4 to 1. The original keyring is visibly moved from drawer 1 to 3. The original keyring is visibly moved from drawer 3 to 2. The original keyring is visibly moved from drawer 2 to 1. The original keyring is visibly moved from drawer 1 to 3. The original keyring is visibly moved from drawer 3 to 1. The original keyring is visibly moved from drawer 1 to 3. Therefore the answer is Drawer 3.

K.5 Example 05: Object ownership across exchanges

Original scene 033; Original development; 50 s; assigned level 5 (not human-calibrated).



Puzzle. Who has the screwdriver that began with worker 3?

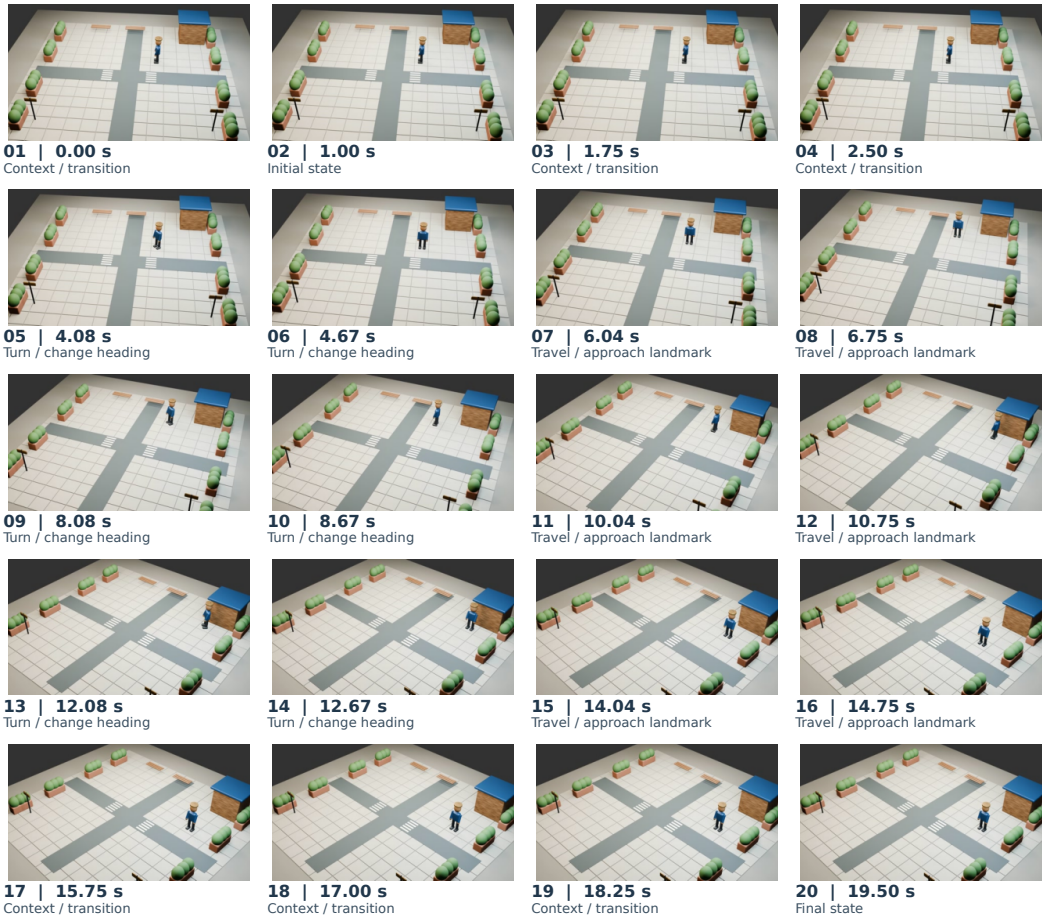
Track the tool initially held by worker 3. The other tool looks the same. Shirt numbers identify the workers.

A. Worker 4
B. Worker 1
C. Worker 2
D. Worker 3

Correct answer: A (Worker 4). The original screwdriver is visibly moved from worker 3 to 2. The original screwdriver is visibly moved from worker 2 to 1. The original screwdriver is visibly moved from worker 1 to 2. The original screwdriver is visibly moved from worker 2 to 1. The original screwdriver is visibly moved from worker 1 to 4. The original screwdriver is visibly moved from worker 4 to 3. The original screwdriver is visibly moved from worker 3 to 1. The original screwdriver is visibly moved from worker 1 to 4. Therefore the answer is Worker 4.

K.6 Example 06: Crossing paths and hidden motion

Original scene 046; Original development; 20 s; assigned level 2 (not human-calibrated).



Puzzle. Relative to the courier's final forward direction, which way was the courier facing at the start?

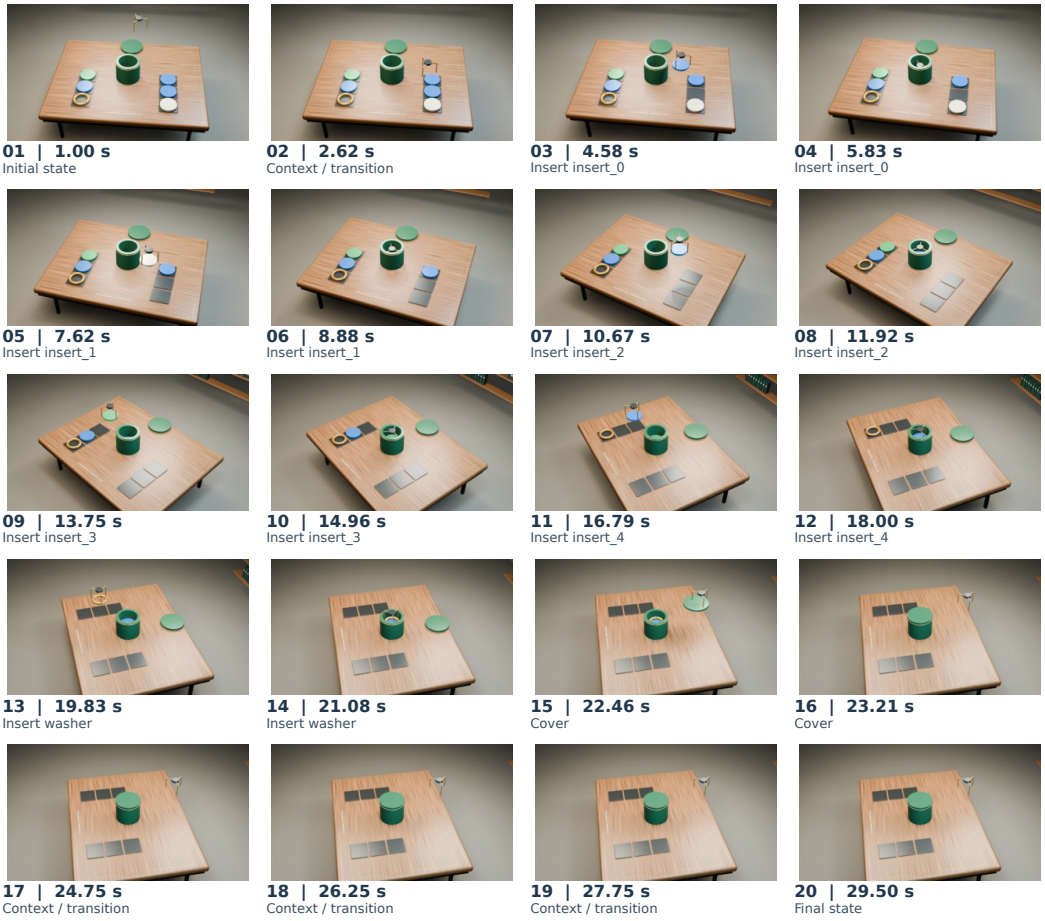
Track the courier's movements and gold forward arrow. The camera also moves. Signs and other stationary arrows do not control the route.

A. Directly behind
B. To the right
C. To the left
D. Straight ahead

Correct answer: B (To the right). Turn left through 90 degrees. Turn right through 90 degrees. Turn left through 90 degrees. Therefore the answer is To the right.

K.7 Example 07: Remembering assembly order

Original scene 065; Original development; 30 s; assigned level 3 (not human-calibrated).



Puzzle. Which colour is two layers below the brass washer?

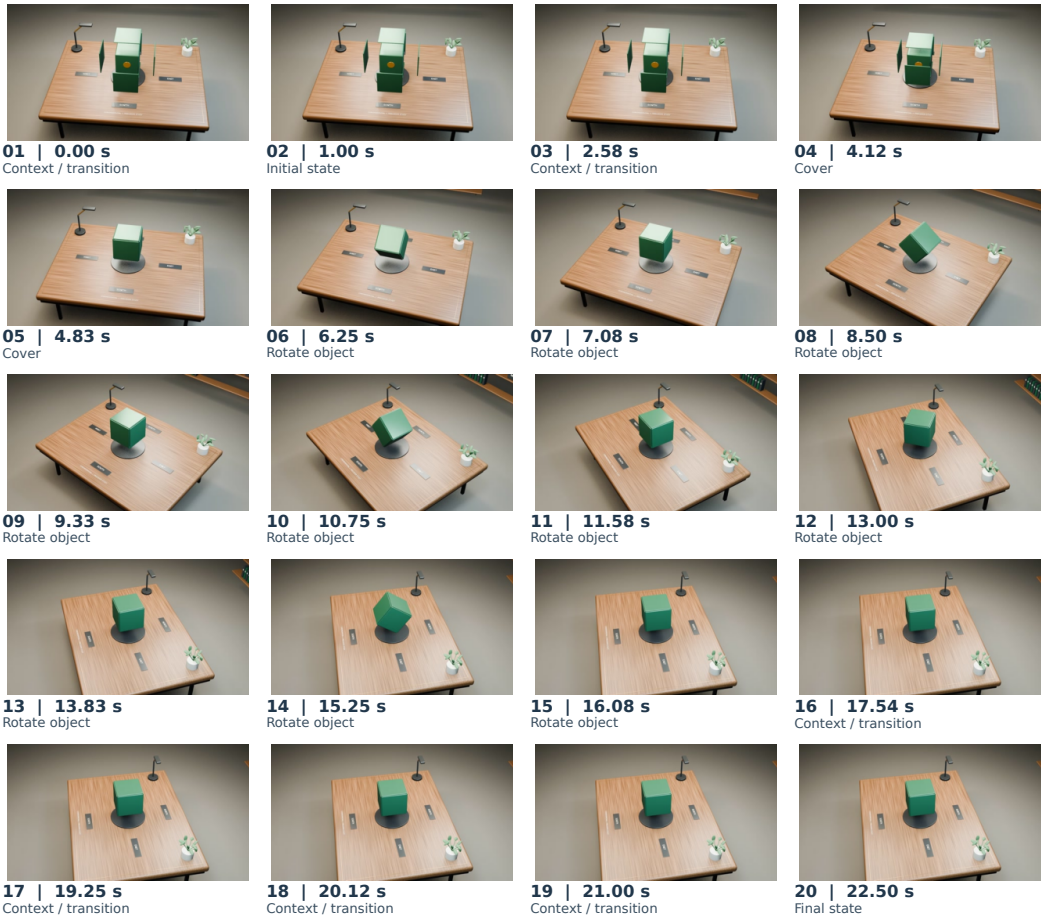
Remember the insert colours and their order. Each part rests on the previous one. The opaque housing prevents seeing the full stack after assembly.

A. Green insert
B. Ivory insert
C. Coral insert
D. Blue insert

Correct answer: A (Green insert). Insert the blue disk as layer 1, lowering it visibly into the opaque cartridge. Insert the ivory disk as layer 2, lowering it visibly into the opaque cartridge. Insert the blue disk as layer 3, lowering it visibly into the opaque cartridge. Insert the green disk as layer 4, lowering it visibly into the opaque cartridge. Insert the blue disk as layer 5, lowering it visibly into the opaque cartridge. Insert the brass washer on top of all previous disks. Therefore the answer is Green insert.

K.8 Example 08: Tracking a face through 3D rotations

Original scene 076; Original development; 23 s; assigned level 3 (not human-calibrated).



Puzzle. Which world direction does the face with the original brass dot point toward at the end?

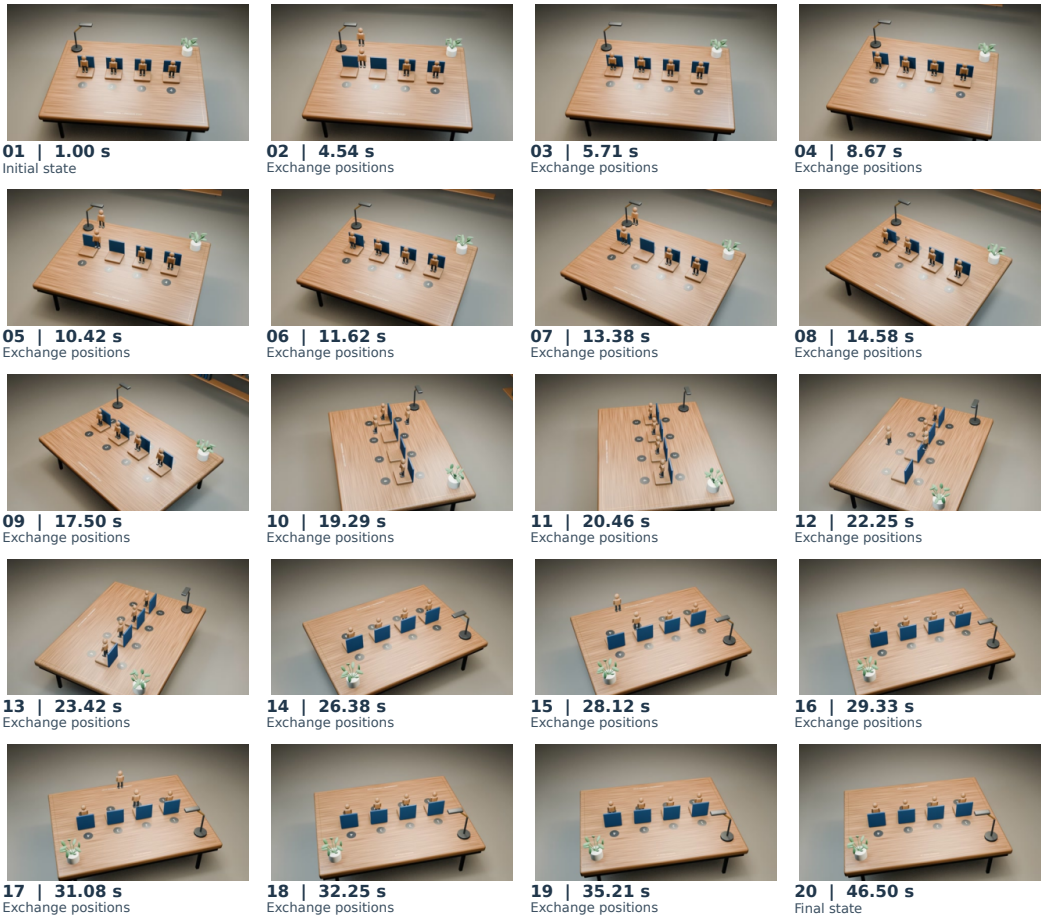
Remember the dotted face before the covers close. The cube rotates around fixed world axes. The compass labels stay on the table; up and down refer to the room.

A. South
B. Down
C. Up
D. North

Correct answer: D (North). All six matching face covers close, hiding the initial brass dot. Rotate the fixture +90 degrees around the fixed world X axis; the concealed face remains attached. Rotate the fixture +90 degrees around the fixed world Y axis; the concealed face remains attached. Rotate the fixture +90 degrees around the fixed world Y axis; the concealed face remains attached. Rotate the fixture -90 degrees around the fixed world Z axis; the concealed face remains attached. Rotate the fixture -90 degrees around the fixed world X axis; the concealed face remains attached. Therefore the answer is North.

K.9 Example 09: Identity through queue rearrangements

Original scene 090; Original development; 47 s; assigned level 5 (not human-calibrated).



Puzzle. At the end, which numbered place holds the token that started at place 1?

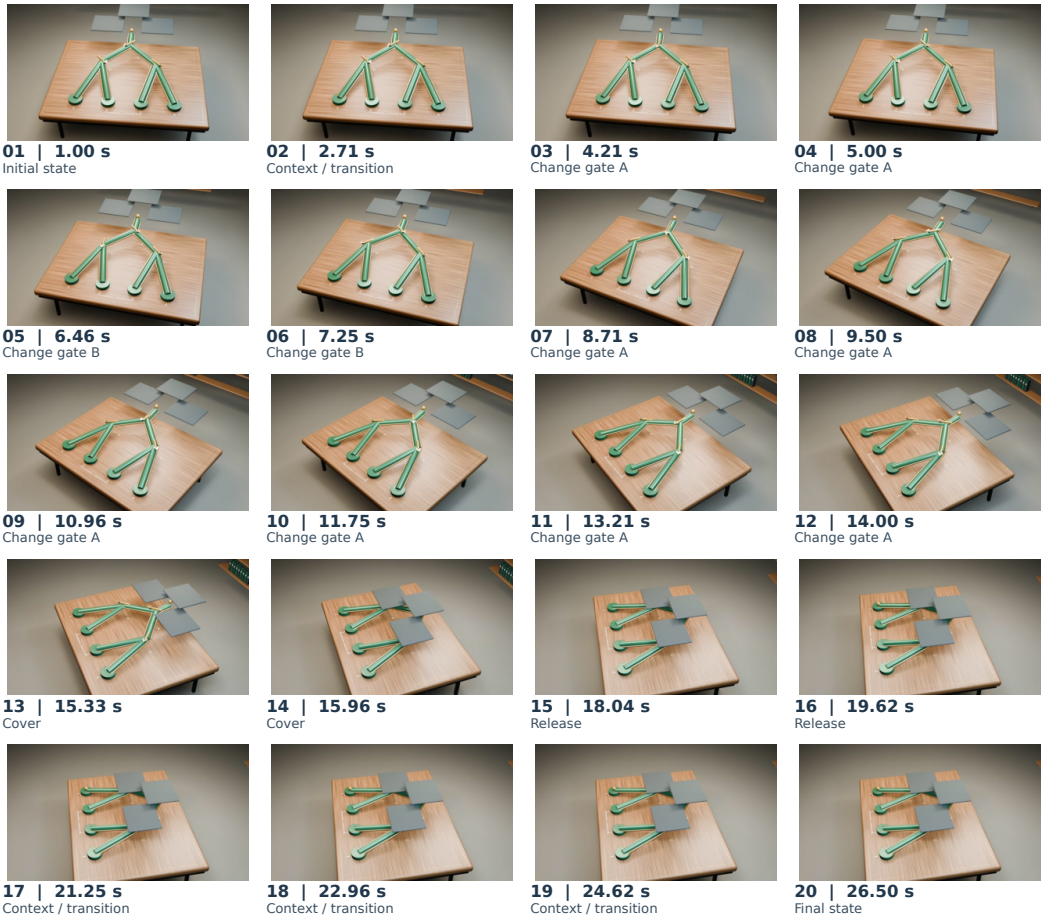
Track the identical wooden token that begins at place 1. The numbered places stay fixed; the tokens exchange positions.

A. Position 4
B. Position 2
C. Position 1
D. Position 3

Correct answer: B (Position 2). Replaying the 11 scripted events with persistent identities gives Position 2. The images show action boundaries and intervening motion; the stored answer is independently checked by the renderer-independent event interpreter.

K.10 Example 10: Remembering a mechanical configuration

Original scene 095; Original development; 27 s; assigned level 3 (not human-calibrated).



Puzzle. If the last move of lever A had not happened, which numbered catcher would receive the marble?

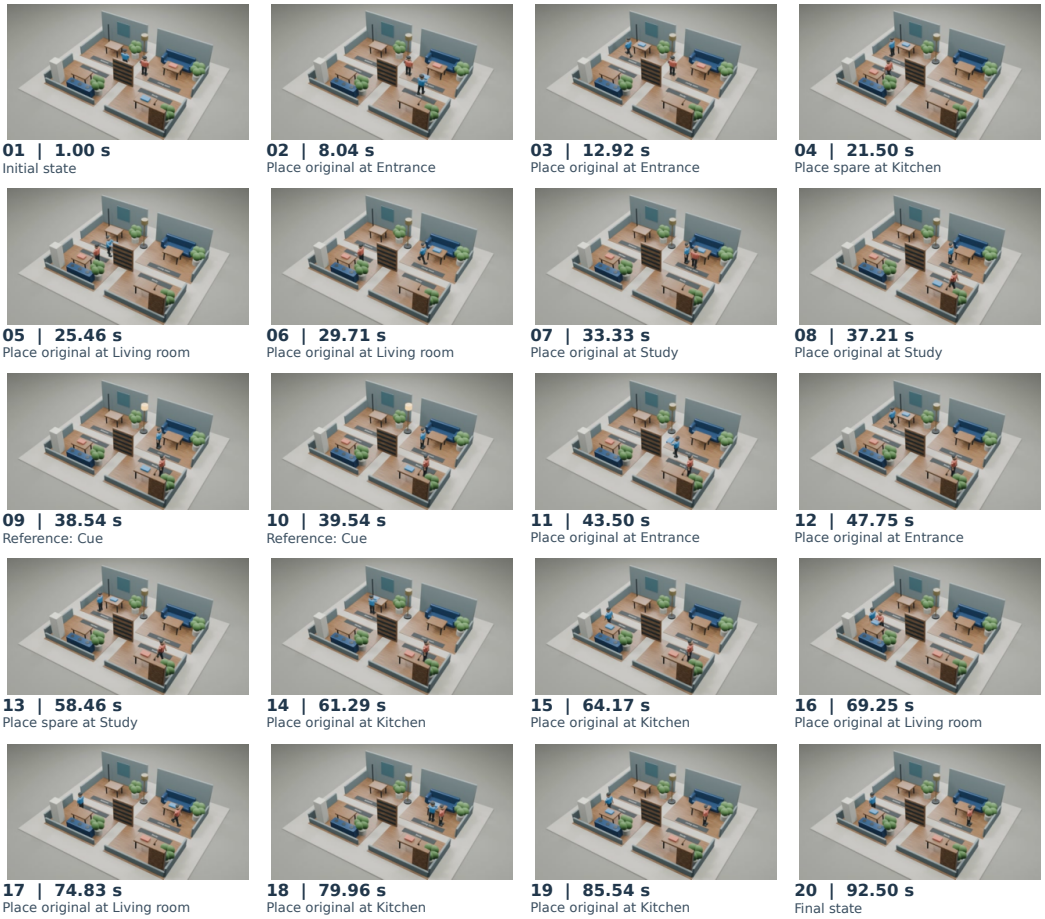
Track levers A, B and C before the screen closes. A selects the left or right branch; B selects 1 or 2; C selects 3 or 4. Imagine omitting only the last move of A.

A. Catcher 2
B. Catcher 3
C. Catcher 4
D. Catcher 1

Correct answer: C (Catcher 4). Move lever A to its opposite detent, switching which branch is open. Move lever B to its opposite detent, switching which branch is open. Move lever A to its opposite detent, switching which branch is open. Move lever A to its opposite detent, switching which branch is open. Move lever A to its opposite detent, switching which branch is open. Omit the last move of lever A, keeping every other lever change unchanged. Therefore the answer is Catcher 4.

K.11 Example 11: A past location at a light cue

Original scene 103; Original evaluation; illustrated after result freeze; 93 s; assigned level 5 (not human-calibrated).



Puzzle. When the standing lamp first lit amber, where was the blue book?

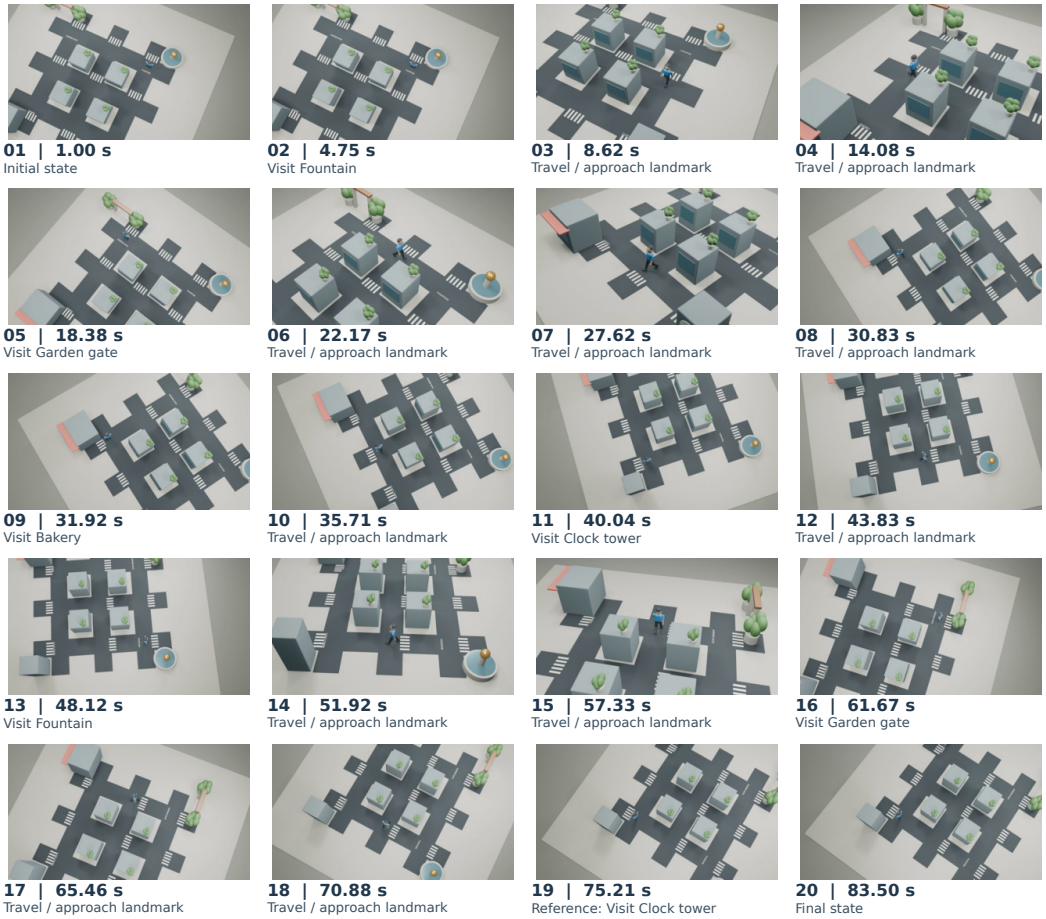
Remember the earlier lamp moment while following the people and their deliveries. Follow the blue book; the coral book is a distractor. Pickups and placements are visible; objects stay with the carrier while out of sight.

A. Living room
B. Kitchen
C. Study
D. Entrance

Correct answer: C (Study). The tracked item followed this sequence before the lamp lit: Study → Entrance → Living room → Study. Later deliveries do not change its location at that earlier moment.

K.12 Example 12: Landmark history with moving viewpoints

Original scene 201; Original evaluation; illustrated after result freeze; 84 s; assigned level 5 (not human-calibrated).



Puzzle. Which landmark did Mira visit 3 landmark visits before her final stop at the Clock tower?

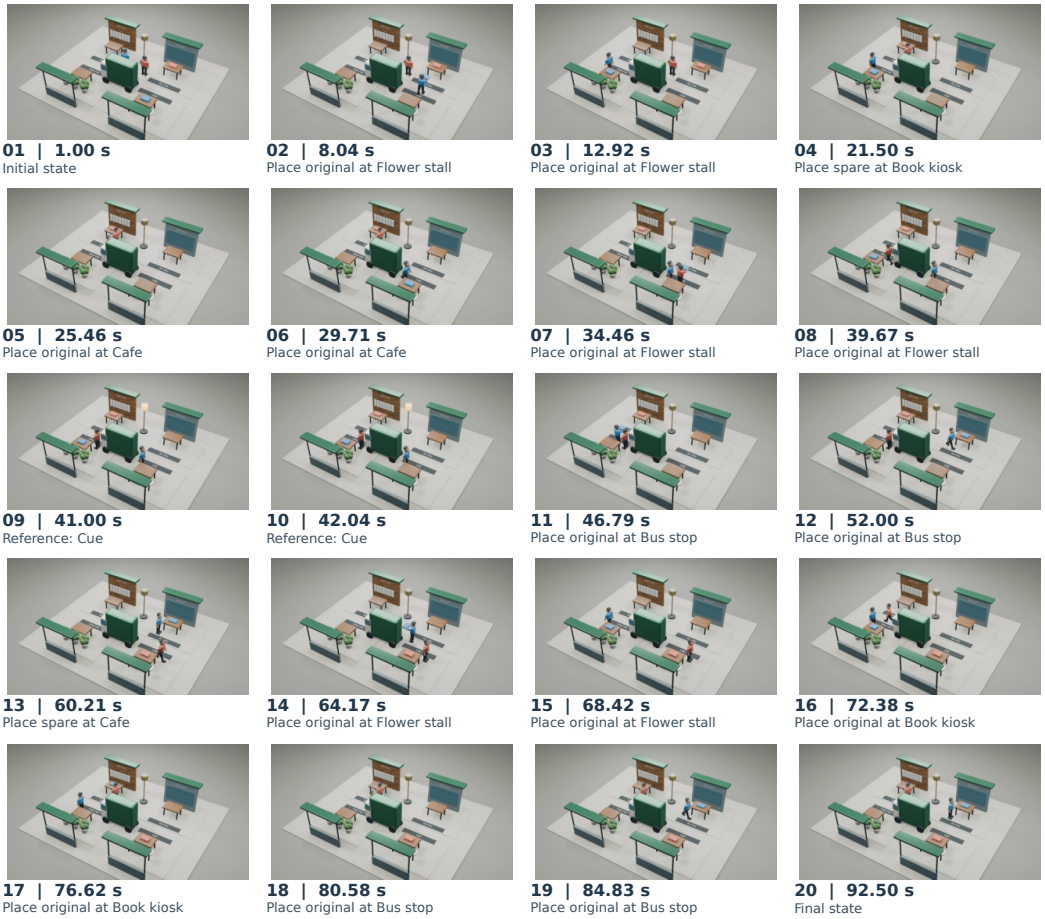
Follow Mira through the streets. A landmark visit means she stops beside one of the four named landmarks. Repeated visits count again. Remember the visit order through the turns and brief occlusions. The camera also changes viewpoint; follow Mira and the fixed landmarks as the view rotates.

A. Clock tower
B. Garden gate
C. Fountain
D. Bakery

Correct answer: A (Clock tower). The landmark visits up to the indicated moment were: Fountain → Garden gate → Bakery → Clock tower → Fountain → Garden gate → Clock tower. Counting back 3 landmark visits gives Clock tower.

K.13 Example 13: Deliveries behind an occluder

Original scene 161; Original evaluation; illustrated after result freeze; 93 s; assigned level 5 (not human-calibrated).



Puzzle. When the standing lamp first lit amber, where was the blue book?

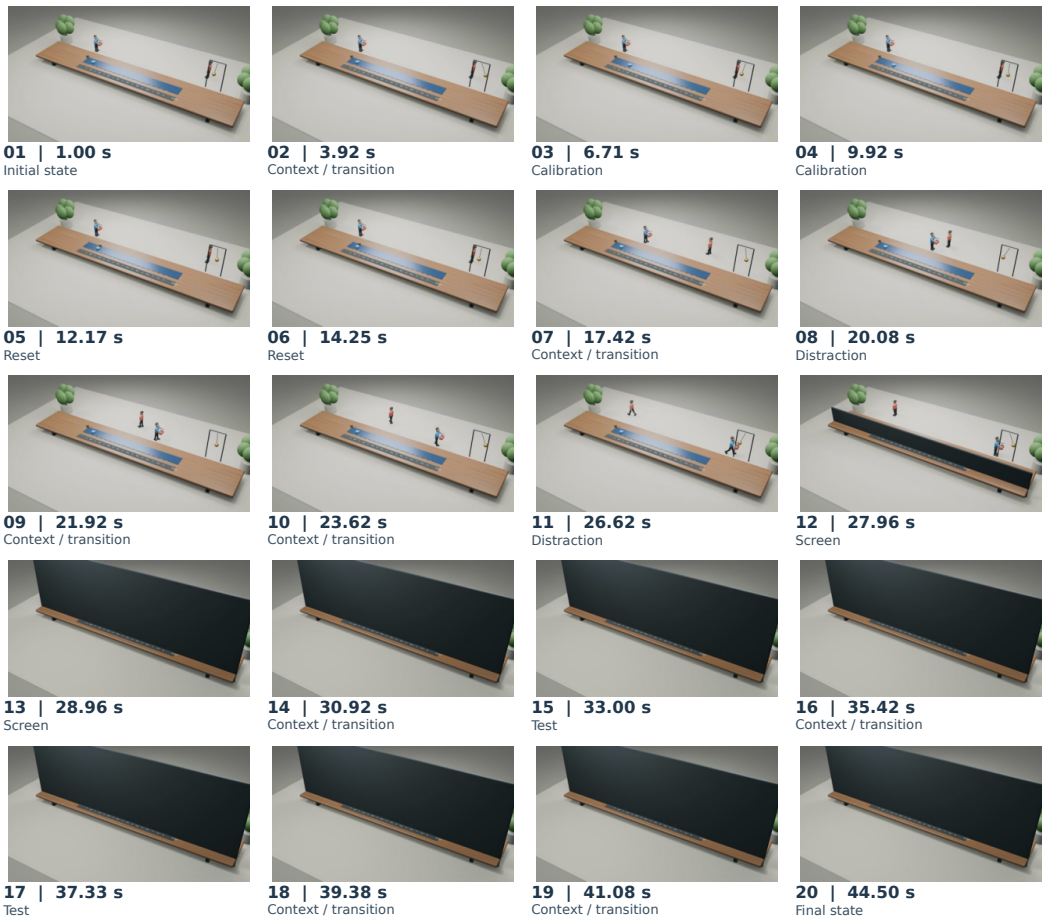
Remember the earlier lamp moment while following the people and their deliveries. Follow the blue book; the coral book is a distractor. Pickups and placements are visible; objects stay with the carrier while out of sight.

A. Flower stall
B. Bus stop
C. Book kiosk
D. Cafe

Correct answer: A (Flower stall). The tracked item followed this sequence before the lamp lit: Cafe → Flower stall → Cafe → Flower stall. Later deliveries do not change its location at that earlier moment.

K.14 Example 14: Remembering a physical calibration

Original scene 181; Original evaluation; illustrated after result freeze; 45 s; assigned level 5 (not human-calibrated).



Puzzle. After the delayed launch at twice the calibration speed, how far does the puck travel before stopping?

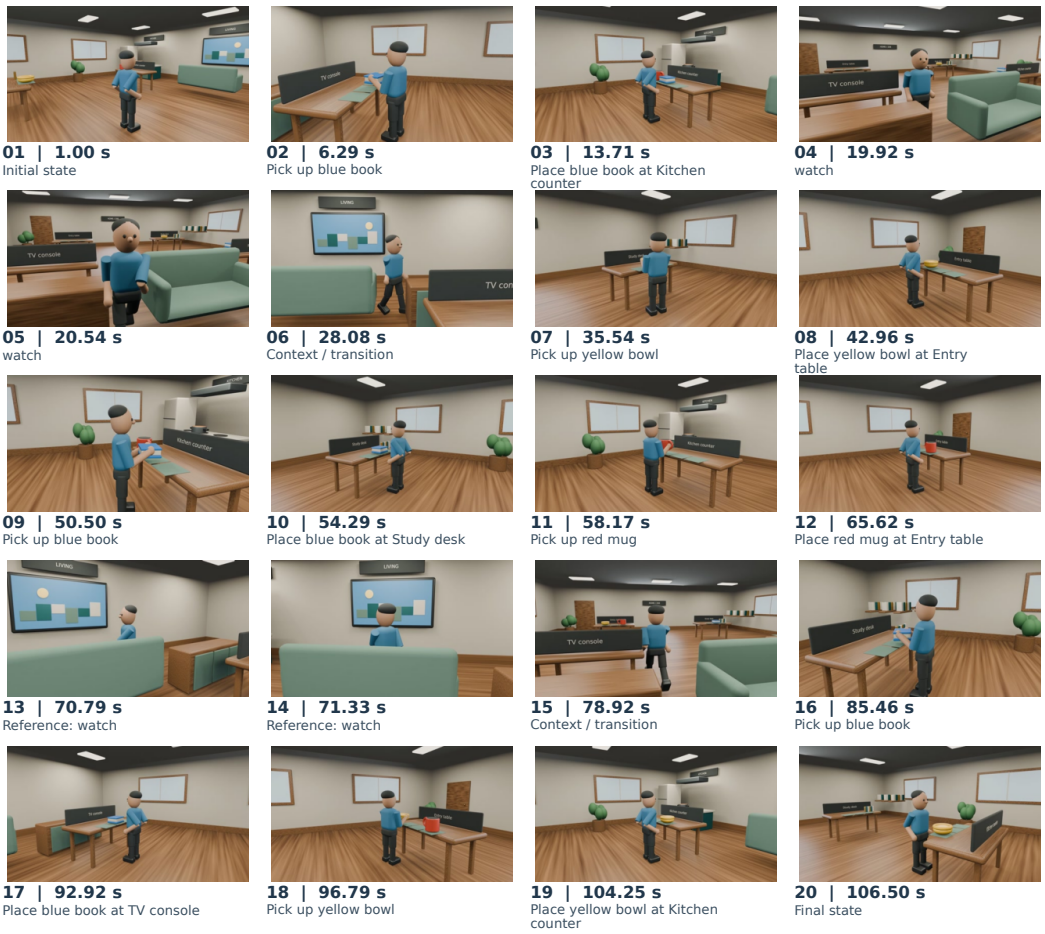
Remember where the blue puck stopped in the first trial. The second launch is exactly twice as fast on the same horizontal surface. The puck slides without rolling; friction stays constant.

A. 4 metres
B. 6 metres
C. 8 metres
D. 2 metres

Correct answer: A (4 metres). The calibration stopped after 1 m. With the same friction, stopping distance is proportional to speed squared. Doubling speed multiplies the distance by four: 4 m.

K.15 Example 15: Activities inside a roofed home

Original scene 208; Original evaluation; illustrated after result freeze; 107 s; assigned level 5 (not human-calibrated).



Puzzle. Where was the blue book when the person sat down to watch TV for the SECOND time?

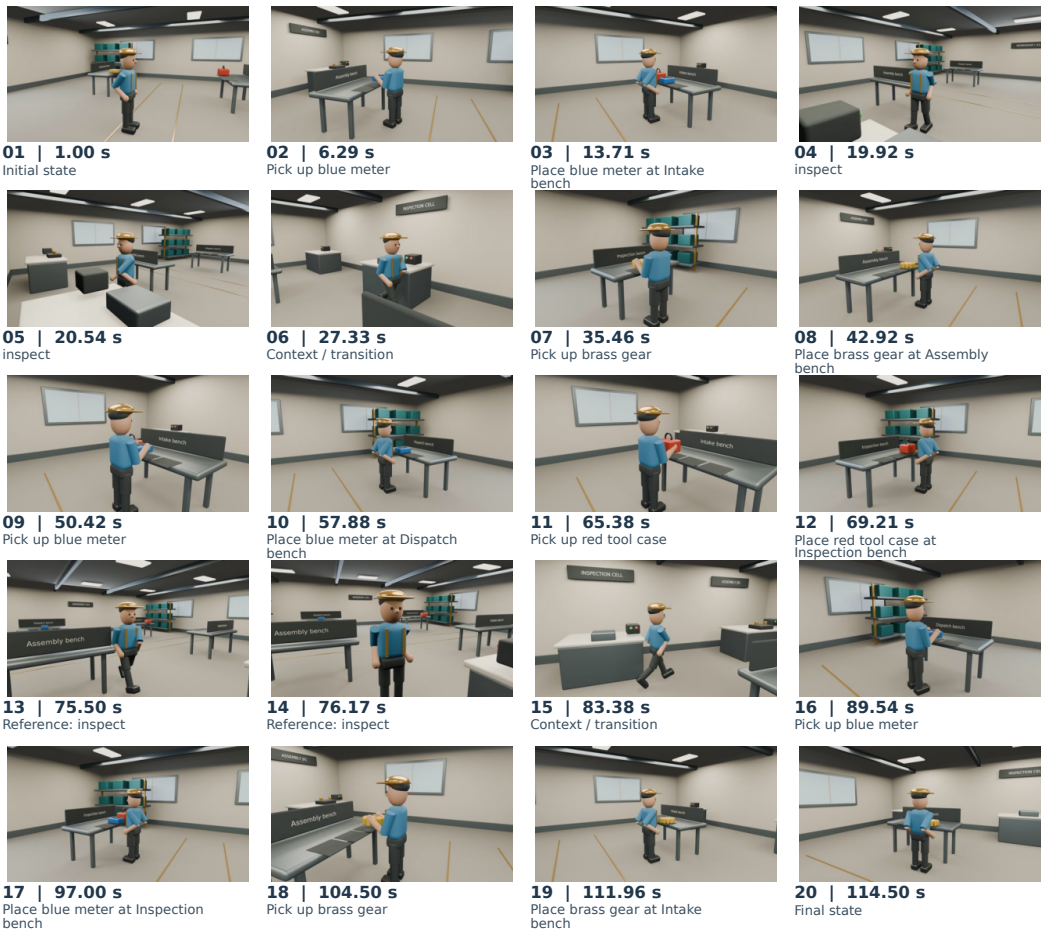
Follow the three distinct objects as the person moves them. Remember the order and earlier locations while the camera moves around the person. Activities may repeat. Answer after watching; the explanation stays hidden until you choose.

A. Study desk
B. Entry table
C. Kitchen counter
D. TV console

Correct answer: A (Study desk). Before the specified activity, the blue book followed: TV console → Kitchen counter → Study desk. Its location at the requested activity was Study desk.

K.16 Example 16: Item history inside a workshop

Original scene 213; Original evaluation; illustrated after result freeze; 115 s; assigned level 5 (not human-calibrated).



Puzzle. Where was the blue meter at the START of the SECOND inspection of the fixed test block?

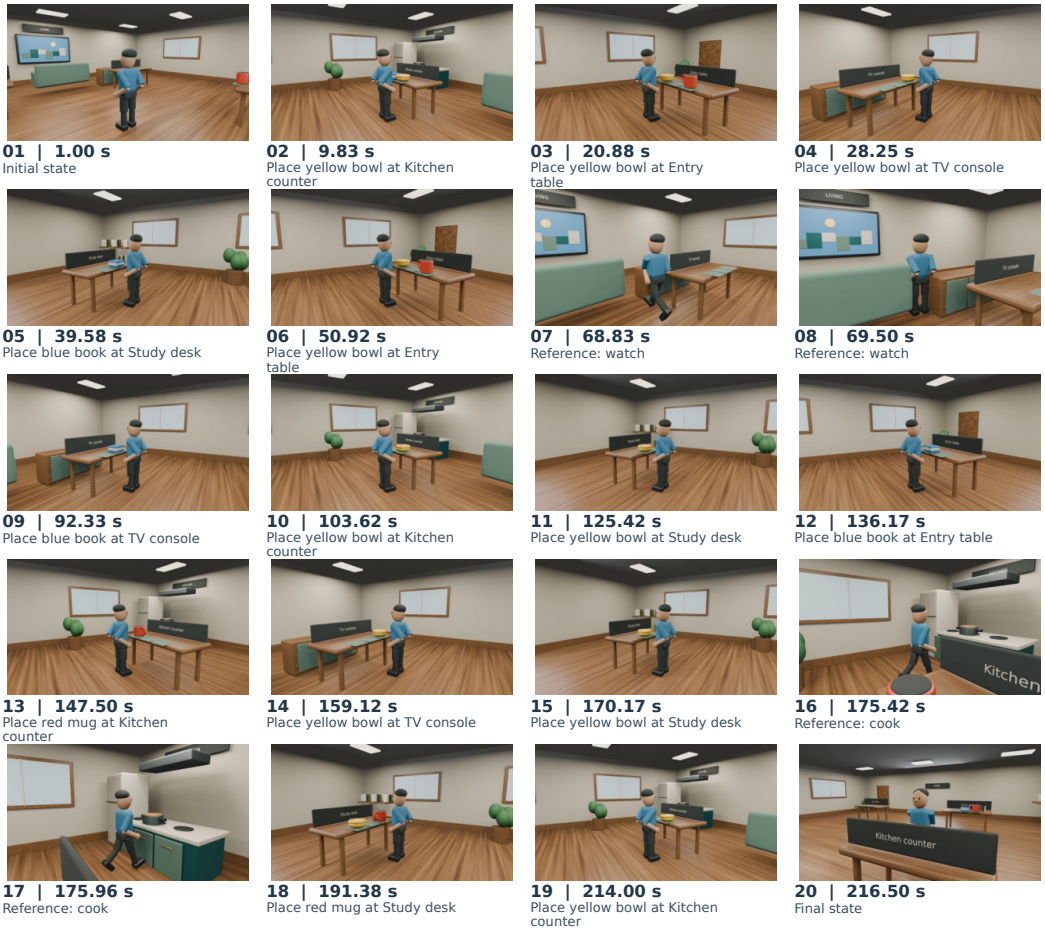
Follow the three distinct objects as the person moves them. Remember the order and earlier locations while the camera moves around the person. Activities may repeat. Answer after watching; the explanation stays hidden until you choose.

A. Dispatch bench
B. Intake bench
C. Assembly bench
D. Inspection bench

Correct answer: A (Dispatch bench). Before the specified activity, the blue meter followed: Assembly bench → Intake bench → Dispatch bench. Its location at the requested activity was Dispatch bench.

K.17 Example 17: Binding repeated activities to object history

Extension l2 tuning 00; Extension development; 217 s; assigned level 2 (not human-calibrated).



Puzzle. Consider the item in the most recent delivery to the study desk before the second cooking activity began. Where was that item when the first television break began?

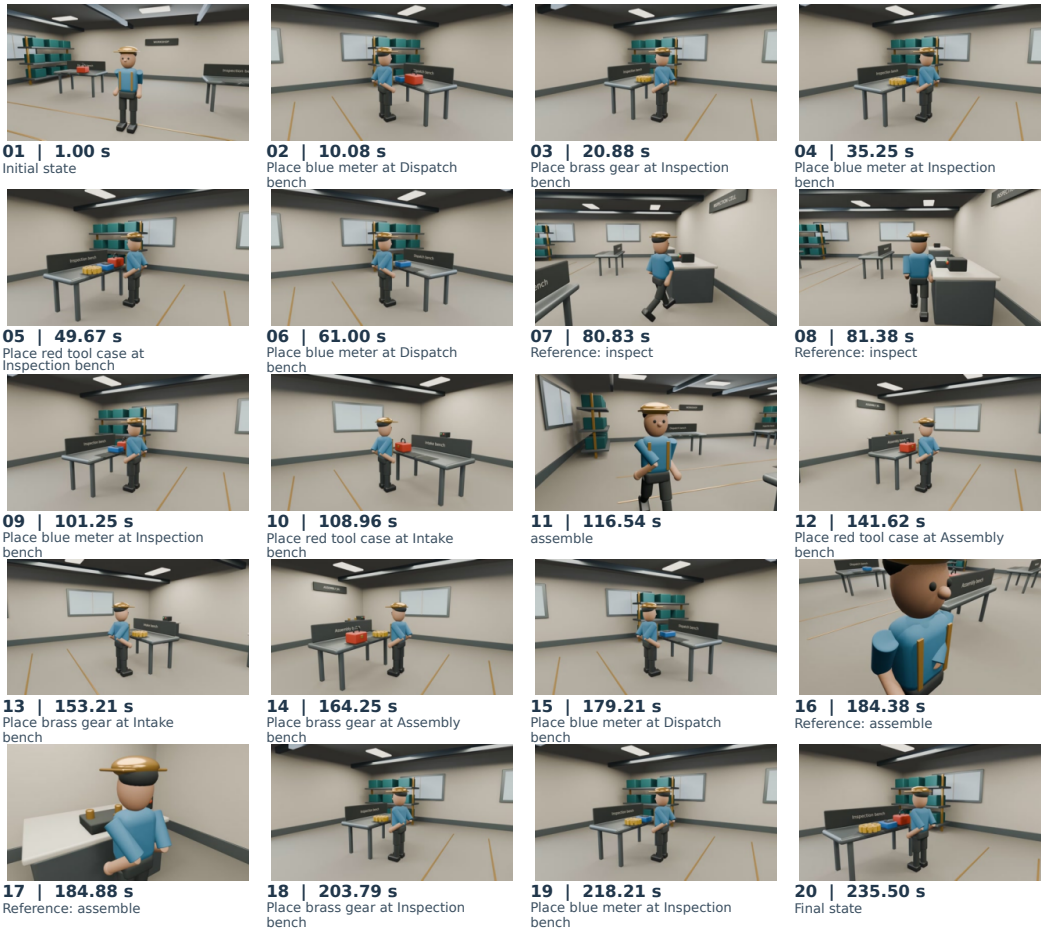
Watch the person move three distinguishable items between four named stations. A delivery is a completed placement; repeated deliveries of the same item count separately. Locations refer to the fixed stations. Activities may repeat.

A. TV console
B. Entry table
C. Study desk
D. Kitchen counter

Correct answer: B (Entry table). Delivery rank 1 selects the yellow bowl (placement completed at 170.6 s). Before the earlier activity at 63.0 s, its locations were: Study desk → Kitchen counter → Entry table → TV console → Entry table. Counting back 0 completed relocations gives Entry table.

K.18 Example 18: Binding workshop deliveries to earlier states

Extension l2 tuning 01; Extension development; 236 s; assigned level 2 (not human-calibrated).



Puzzle. Consider the item in the most recent delivery to the dispatch bench before the second assembly activity began. Where was that item when the first inspection activity began?

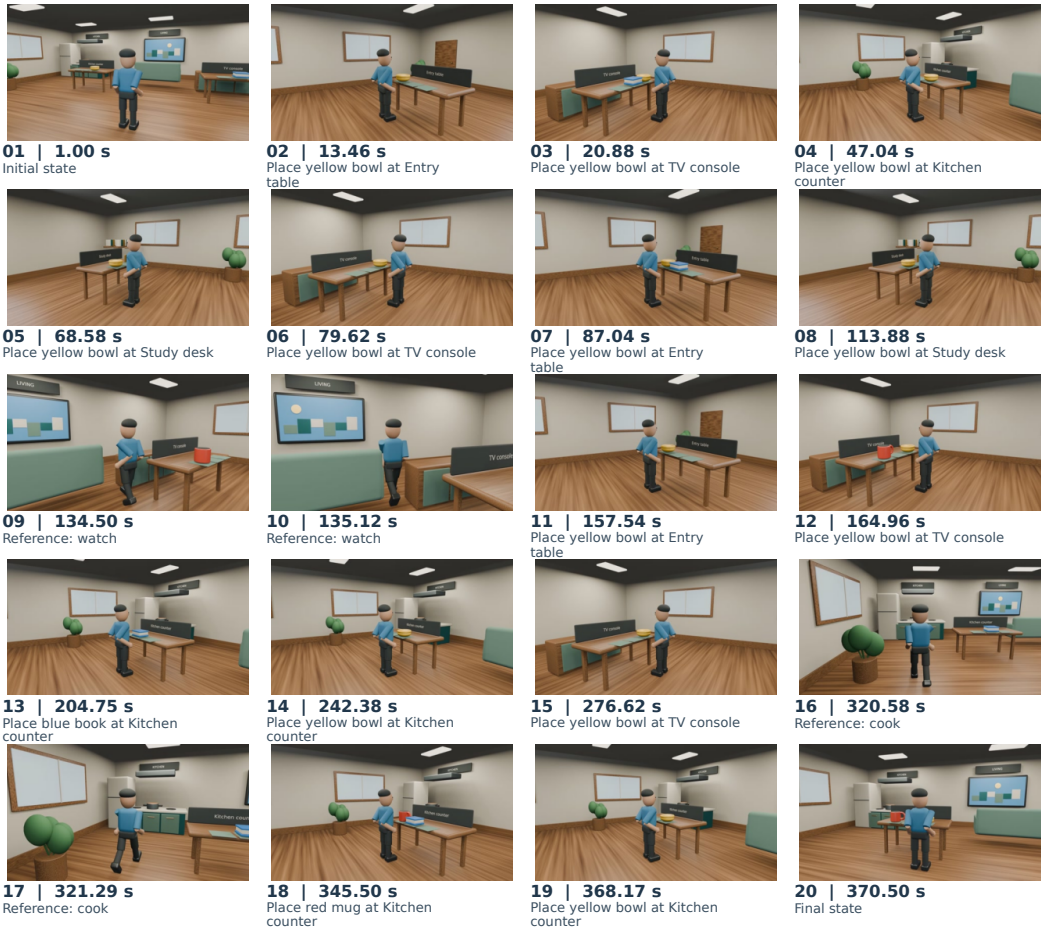
Watch the person move three distinguishable items between four named stations. A delivery is a completed placement; repeated deliveries of the same item count separately. Locations refer to the fixed stations. Activities may repeat.

A. Assembly bench
B. Inspection bench
C. Intake bench
D. Dispatch bench

Correct answer: D (Dispatch bench). Delivery rank 1 selects the blue meter (placement completed at 179.7 s). Before the earlier activity at 75.8 s, its locations were: Assembly bench → Dispatch bench → Inspection bench → Dispatch bench. Counting back 0 completed relocations gives Dispatch bench.

K.19 Example 19: Binding repeated activities to object history

Extension l6 tuning 00; Extension development; 371 s; assigned level 6 (not human-calibrated).



Puzzle. Consider the item in the third-most-recent delivery to the tv console before the second cooking activity began. At the start of the second television break, which station was 2 completed relocations earlier in that item's location history than its location at that moment?

Watch the person move three distinguishable items between four named stations. A delivery is a completed placement; repeated deliveries of the same item count separately. Locations refer to the fixed stations. Activities may repeat.

A. Kitchen counter
B. Study desk
C. TV console
D. Entry table

Correct answer: C (TV console). Delivery rank 3 selects the yellow bowl (placement completed at 277.1 s). Before the earlier activity at 128.7 s, its locations were: Kitchen counter → Entry table → TV console → Kitchen counter → Study desk → TV console → Entry table → Study desk. Counting back 2 completed relocations gives TV console.

K.20 Example 20: Binding workshop deliveries to earlier states

Extension l6 tuning 01; Extension development; 345 s; assigned level 6 (not human-calibrated).



Puzzle. Consider the item in the third-most-recent delivery to the assembly bench before the second assembly activity began. At the start of the second inspection activity, which station was 2 completed relocations earlier in that item's location history than its location at that moment?

Watch the person move three distinguishable items between four named stations. A delivery is a completed placement; repeated deliveries of the same item count separately. Locations refer to the fixed stations. Activities may repeat.

A. Dispatch bench
B. Inspection bench
C. Assembly bench
D. Intake bench

Correct answer: A (Dispatch bench). Delivery rank 3 selects the brass gear (placement completed at 273.1 s). Before the earlier activity at 127.0 s, its locations were: Intake bench → Dispatch bench → Assembly bench → Inspection bench → Intake bench → Assembly bench → Dispatch bench → Intake bench → Assembly bench. Counting back 2 completed relocations gives Dispatch bench.