TRACER: A MULTIMODAL HARNESS FOR PATIENT CARE IN HOSPITALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous video and clinical records provide complementary information for understanding how patient behavior evolves over a hospital stay. A central challenge is how to represent behavior across the entire stay while preserving access to fine-grained visual and clinical evidence. We introduce TRACER, a training-free multimodal harness that constructs minute-level action records from video and combines them with clinical records into hourly reports. The harness equips an LLM agent with tools to query this hierarchy, inspect relevant video, and retrieve clinical records. To evaluate TRACER, we construct two benchmarks. The first contains 397 annotated one-minute clips spanning 15 atomic actions; the second poses 144 clinician-informed multiple-choice questions, grounded in 264 hours of video and clinical records from two ICU patients. TRACER achieves up to 84.8% accuracy for action detection and 61.0% QA accuracy versus 49.3% for the strongest baseline using the same backbone. Across 280 patients spanning approximately 60,000 camera-hours, TRACER identifies significantly less purposeful activity in patients with a positive delirium assessment during their stay than in those without ($p < 0.0001$). Our results show that combining continuous video with clinical records can support understanding of patient behavior over time and across patient cohorts without training on protected health data.

1 INTRODUCTION

Reasoning over long, continuous streams of visual observations is important for understanding human behavior. Recent advances in long-form video understanding have improved vision-language models' (VLMs) ability to recognize, summarize, and reason about events and activities over increasingly long visual contexts (Chandrasegaran et al., 2024; Fu et al., 2024; Wu et al., 2024; Zhang et al., 2025; Yang et al., 2025). However, understanding real-world human behavior often requires integrating visual observations with external information. This need is particularly important in healthcare, where behavioral changes provide evidence about a patient's condition but must be interpreted within the context of their medical history and ongoing treatment. In this work, we study this problem in the intensive care unit (ICU), a high-stakes environment where continuous video and clinical records together capture changes in patient movement, responsiveness, interaction, and care across days (Devlin et al., 2018).

Developing this capability in the ICU presents several technical challenges. Long-context vision–language models and caption-based pipelines offer approaches for long-form video understanding (Chandrasegaran et al., 2024; Zeng et al., 2023), but applying them across entire ICU stays requires balancing temporal coverage with access to detail. Specifically, a system must be able to connect brief movements and interactions to behavioral patterns extending across days. And while bedside video captures observable activity and interactions, electronic health records (EHRs) document clinical assessments and interventions that provide complementary behavioral context. Finally, acquiring continuous bedside recordings linked to EHRs requires dedicated sensing infrastructure and secure handling of sensitive patient information; the recordings and clinical records used in our study cannot be released publicly. We therefore seek a method for longitudinal understanding that can be shared without releasing protected patient records or model parameters fitted to them.

To address these challenges, we introduce TRACER (Trajectory Reasoning Across Clinical Event Records), a training-free multimodal harness for understanding multi-day ICU patient trajectories.

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.

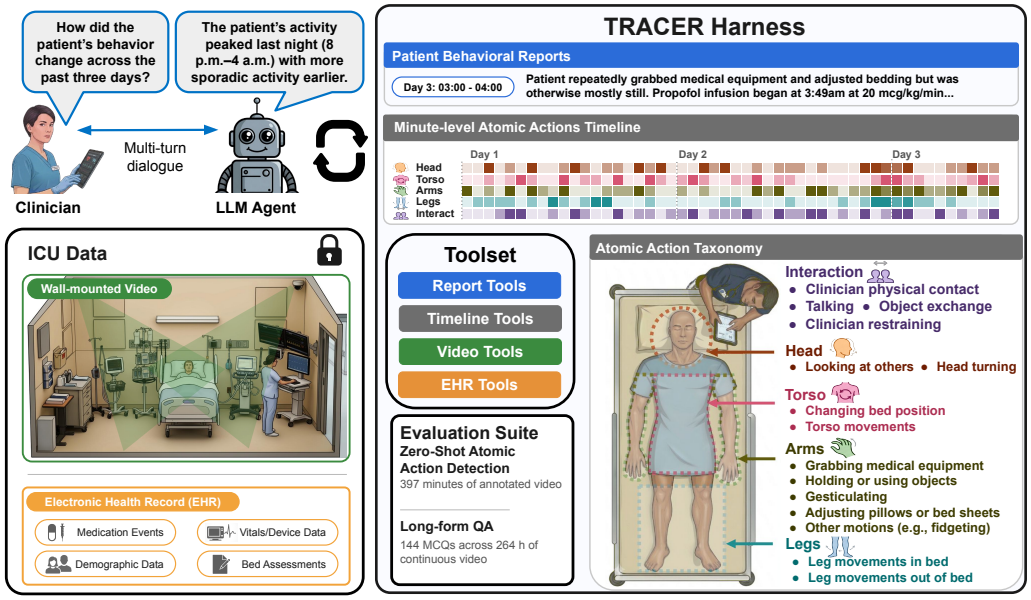


Figure 1: **Overview of TRACER.** TRACER equips an agent model with hierarchical behavioral representations and tools for understanding multi-day ICU patient trajectories. Vision–language models analyze one-minute bedside video segments to detect atomic actions defined by a clinically informed taxonomy and generate descriptive captions. These outputs are combined with concurrent medication records to generate structured hourly behavioral reports. Through report, timeline, video, and EHR tools, the agent iteratively retrieves evidence and examines relevant video segments to answer retrospective clinical questions. Our evaluation suite assesses atomic action detection on 397 annotated one-minute clips and longitudinal clinical reasoning on 144 multiple-choice questions spanning 264 hours of continuous video.

TRACER separates temporal abstraction, which builds a hierarchy of minute- and hour-level records continuously as the stay is recorded, from evidence acquisition, in which an agent reasons over these compact records and accesses the raw video and EHR as needed. To construct this hierarchy, vision–language models (VLMs) analyze one-minute video segments zero-shot, producing atomic action detections under a clinically informed taxonomy together with descriptive captions. A language model combines these outputs with concurrent medication records to generate structured hourly behavioral reports. The atomic action records supply consistent variables for aggregation and comparison, while captions and reports provide descriptive context beyond the predefined categories. A tool-equipped agent model can then retrieve reports, compare action predictions across time, query clinical records, and request targeted visual analysis of selected video segments, enabling retrospective investigation (see Figure 1).

We evaluate TRACER using recordings and linked EHR data collected over four years at an academic medical center. We examine the accuracy of local perception, the use of these observations for longitudinal reasoning, and the clinical relevance of cohort-level behavioral predictions. On our benchmark of 397 manually annotated one-minute clips from 26 patient stays, TRACER achieves up to 84.8% accuracy and 68.6 micro-F1 across 15 atomic action categories. To probe reasoning over patient trajectories, we introduce TRAJECTORY-144, a clinician-informed benchmark of 144 multiple-choice questions grounded in video and clinical records from two patient stays spanning 264 hours of continuous video and EHR data. TRACER reaches 76.9% QA accuracy with the best proprietary LLM agent and 67.9% QA accuracy for the best open-weight LLM agent, while no long-context VLM or Socratic caption baseline (Zeng et al., 2023; Chandrasegaran et al., 2024) exceeds 50%. Finally, we apply TRACER across a cohort of 280 patients spanning approximately 60,000 camera-hours and detect significantly lower frequencies of purposeful activities among patients with a positive delirium assessment during their stay, indicating that TRACER can quantify clinically relevant behavioral differences between patient groups.

Table 1: **MCQ performance on TRAJECTORY-144.** QA accuracy (%) is shown by question type and is averaged over 5 gold-answer placements. TRACER rows share an identical harness (prompts, taxonomy, and temporal hierarchy) and differ only in the agent model. Best per column is in bold, and second best is underlined. The video tool is served by the agent model, or by Gemini 3.1 Flash Lite where marked †.

	Video	EHR	Cross-modal	Avg.
No video				
Random	20.1	20.0	20.0	20.1
Text-only LLM, Gemini 3.1 Flash Lite	32.3	26.9	25.0	30.9
Long-context VLMs				
Qwen3-VL 8B, 475 frames	31.6	10.0	47.5	28.6
Qwen3-VL 8B, 950 frames	36.3	13.8	50.0	33.0
Gemini 3.1 Flash Lite, 950 frames	36.5	14.6	25.0	31.9
Gemini 3.1 Flash Lite, 475 frames + EHR	41.8	83.8	40.0	49.3
Socratic models (HourVideo)				
Qwen3-VL 8B	33.7	22.3	20.0	30.9
Qwen3.5 9B	30.4	18.5	10.0	27.1
Gemini 3.1 Flash Lite	36.8	16.2	47.5	33.7
Gemini 3.1 Flash Lite + EHR	37.8	76.2	42.5	45.0
TRACER harness (ours)				
w/ gpt-oss-20b†	50.3	93.1	40.0	57.4
w/ Qwen3.5 9B	49.5	95.4	<u>52.5</u>	57.9
w/ Gemini 3.1 Flash Lite	53.9	96.2	45.0	61.0
w/ Gemma-4 31B†	<u>61.8</u>	**100.0**	47.5	<u>67.9</u>
w/ Gemini 3.5 Flash	**72.0**	<u>99.2</u>	**72.5**	**76.9**

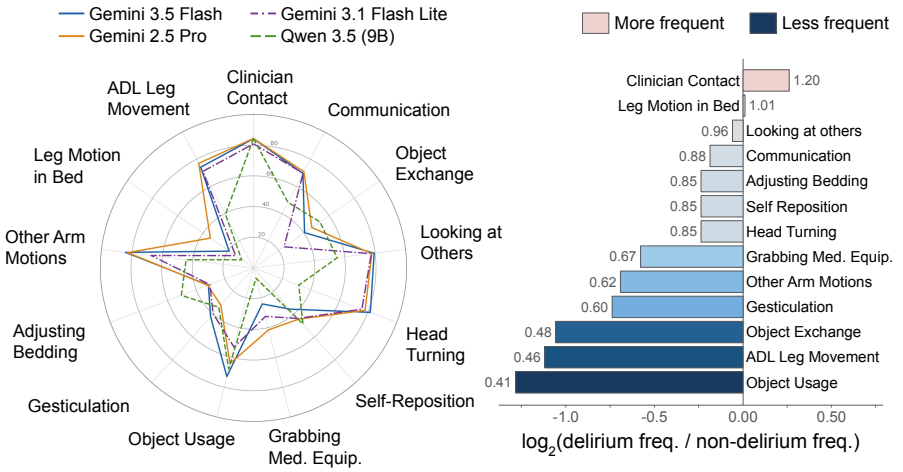
2 Harness Design

Following the terminology of prior work (Packer et al., 2023; Yang et al., 2024; Lee et al., 2026), we call the *harness* everything that surrounds an LLM: the representations it can access, the tools that expose them, and the reasoning loop under which it acts. For our setting (Figure 1), a harness must provide three things: representations of the stay that can capture long-term changes in patient trajectory and moment-level temporal scales, tools that expose those representations together with the raw video and clinical records, and a loop under which a model can gather evidence within a fixed budget.

2.1 Problem Formulation

We model an ICU stay as $\mathcal{S} = \left(\{V_t^{(1)}, V_t^{(2)}\}_{t=1}^{T}, \mathcal{E}\right)$, where $V_t^{(1)}$ and $V_t^{(2)}$ are the one-minute clips from two wall-mounted cameras at minute t, T is the length of the stay in minutes, and $\mathcal{E}$ is the set of timestamped events in the electronic health record (medication administrations, bedside assessments, vital signs, etc.). In our cohort the median stay is 3.3 days ($T \approx 4{,}800$) and stays run up to two weeks ($T \approx 20{,}000$). Given a free-text clinical question q, the task is to return an answer supported by evidence that may lie anywhere in $\mathcal{S}$, in either modality, and at a temporal scale ranging from a moment in time to the entire stay. There are two primary constraints that motivate our system design. First, a single day of video from one camera sampled at 1 FPS is approximately an order of magnitude past the context length of most existing models, which generally range from 256K to 1M tokens (Bai et al., 2025; Comanici et al., 2025) (see Appendix A.6 for more details). Second, a model whose weights are fitted to protected health information cannot be easily shared with external institutions. Thus, we design our harness so that all problem-specific structure and performance gains live in components that can be publicly released (our activity taxonomy, timeline structure and construction, aggregation tools, and model prompts).

TRACER is a harness that separates *temporal abstraction*, which runs continuously in the background as the stay is recorded, from *evidence acquisition*, a process that is performed once per question. During temporal abstraction, as each minute of video arrives it is converted into a minute-level clinical event record consisting of an atomic-action label vector and a caption. As each hour completes, its sixty minute records are fused with the concurrent medication events into an hour-



(a) Minute-level F1 Scores **(b) Cohort Delirium Frequency Ratios**

Figure 2: **Per-Category Breakdown of Atomic Action Performance.** (a) We plot the per-category F1 scores of the atomic actions used by TRACER. (b) We show the relative frequency of predicted atomic actions with respect to delirium status across a cohort of 280 patients spanning $\sim 60,000$ hours of video data. Delirium is defined as any positive CAM-ICU assessment during the stay ($n = 53$ delirium, $n = 227$ non-delirium). Other torso movements and clinician restraining contained too few positive samples to be evaluated accurately and are therefore excluded from both panels.

level report. During evidence acquisition, TRACER supplies a language model agent with a system prompt and a list of tools and accompanying descriptions for examining the hospital stay. Importantly, the tools operate at every level of the hierarchy, including the raw clips V_t and the raw events $\mathcal{E}$. To temporally localize events, the agent uses time cues from the question, from the clinical record, from an action query over the timeline, or from reading through hourly reports (see Table 6 in Appendix A.4 for an exact breakdown). The agent then chooses a tool and its arguments, receives the result in context, and repeats until it commits to an answer or exhausts a fixed budget of calls.

2.2 Harness Architecture

TRACER instantiates this harness with three model roles: a *perception* VLM that turns each minute of video into a clinical event record, a *report* LM that compacts each hour of records and EHR events into a report, and an *agent* model that uses tools and interacts with the user. The first two run continuously as the stay is recorded to build TRACER's temporal hierarchy, allowing the agent to gather information from individual minutes and hourly reports and to view the raw video feed and EHR. The agent model runs per question. No role is trained and each can be replaced (Appendix A.8).

2.3 Minute-Level Clinical Event Records

Taxonomy. The structured representation of a minute is a vector over 15 *atomic actions* in five groups: patient–clinician interaction (contact, talking, object exchange, restraint) and movements of the head, torso, arms, and legs (Figure 1). The taxonomy was developed iteratively with nurses, clinicians, and machine-learning researchers until category definitions were stable. It follows the atomic-action formulation of AVA (Gu et al., 2018) but is chosen for clinical relevance, so that its aggregates can provide information related to mobility, agitation, or clinician attention.

Minute-Level Video Inference. Each one-minute clip from each camera is sampled at 1 fps, and the resulting 60 frames are passed to the perception VLM with a fixed prompt to detect each of the 15 actions and to write a brief free-text caption of the minute (prompt and output schema in Appendix A.3). The two camera views are merged by a logical OR over the 15 labels, resulting in one binary label vector of the 15 actions per patient-minute, alongside the two view-specific captions. This procedure is run over our full cohort of patients, spanning approximately 60,000

camera-hours, enabling the cohort-level analysis of Section 4.3. The choice of perception VLM is primarily a trade-off between accuracy and cost/compute (see Appendix A.9).

2.4 Hourly Behavioral Reports

As each hour of the stay completes, the report model receives its 60 minute-level clinical event records, together with the medication administration events recorded in the EHR during that hour, and produces a structured report with six sections: (1) a movement summary with minute-level citations, (2) a patient–clinician interaction summary, (3) significant moments, (4) medication context, (5) observable activity–medication dynamics, and (6) a brief narrative. We choose hour-level report generation because it matches the most frequent cadence at which entries in ICU flowsheets are typically charted. In order to test the validity of the hourly reports, we audit report faithfulness against the minute records and the medication log (see Appendix A.10).

2.5 Tool-Augmented Trajectory Reasoning

Given a question, the agent model runs a tool-augmented reasoning loop over the TRACER harness. The harness exposes four categories of tools: report, timeline, video, and EHR (exact signatures and the fixed system prompt in Appendix A.2). The timeline and report tools are used to access, aggregate, or search across the temporal hierarchy, while EHR tools can access predefined categories of stored data from the EHR. The video tool enables the agent to perform additional model inference on video data to capture details that may have been missed in the atomic actions and original text caption. Specifically, the video tool allows the agent to author a prompt for a VLM that runs across both camera views (5 fps for Gemini, 2 fps for open-weight models). All output text, tool calls, and tool results are kept in the agent's context. We limit the agent to 25 tool calls, at which point a forced-choice prompt asks it to answer from the evidence gathered so far. The same prompt is used for all evaluated agent models.

3 Data and Evaluation

3.1 ICU Dataset

Continuous video was recorded in the cardiovascular and medical ICUs of a large academic medical center between August 2021 and June 2025 by 16 fixed, corner-mounted cameras covering eight rooms (two per room) at 1920×1080 and 15 fps. The video data was linked to structured EHR data (demographics and encounter metadata, admission–discharge–transfer events, nursing flowsheet data, and the medication administration record). From 2,617 patients assessed for eligibility, sequential clinical and technical exclusion criteria (stay length, sedative exposure, non-neurological admission, room setup, and complete camera coverage) yielded a cohort of 280 patients spanning approximately 60,000 camera-hours. The full selection flow and cohort characteristics are in Appendix A.5. The study was approved by the Institutional Review Board of the study institution.

3.2 Perception Benchmark

To validate minute-level perception we assembled 397 one-minute clips from 26 patient stays, sampled for diversity in time of day, patient acuity, and activity level, and labeled for all 15 atomic actions by one annotator and then verified or corrected by a second ($\kappa = 0.81$, 92.7% agreement on a 20-clip audit; Appendix A.9). We report micro-averaged accuracy, F1, precision, and recall across all 15 atomic actions and 9 perception VLMs in Table 2.

3.3 TRAJECTORY-144: Longitudinal Reasoning Benchmark

Benchmark Construction. To evaluate reasoning over whole stays we created TRAJECTORY-144, a benchmark consisting of 144 five-way multiple-choice questions grounded in 264 hours of continuous video and health records from two ICU stays. To construct the benchmark, each stay was divided into six-hour periods, and an annotator with access to the video and the health records built a timeline of events for that period. Annotators then created questions from these timelines, the record, and the raw video data in consultation with clinicians. Distractors were revised to be plausible and

Table 2: Average atomic action detection performance. Micro-averaged across all 15 atomic actions on the 397-clip test set. Model rows are mean ± std across 5 runs. Sampling settings for every model are listed in Appendix A.8. Best result per column in **bold**.

Model	Accuracy	F1	Precision	Recall
Baselines				
Random (Match distr.)	69.9±0.4	44.8±0.8	44.8±0.8	44.7±1.0
Majority Class	77.7	45.2	68.9	33.6
Open-source				
Qwen3-VL (8B-Instruct)	78.0±0.1	37.6±0.4	**84.8±0.3**	24.2±0.3
Qwen3.5 (4B)	78.2±0.3	41.0±0.8	78.3±1.7	27.7±0.6
Qwen3.5 (9B)	79.7±0.5	48.4±2.0	79.4±1.0	34.8±1.9
Gemma-4 (E4B-it)	74.9±0.1	22.5±0.3	72.2±1.6	13.4±0.2
Proprietary				
Gemini 2.5 Flash Lite	79.0±0.3	53.0±0.7	68.4±0.9	43.3±0.9
Gemini 2.5 Flash	83.2±0.2	66.5±0.5	71.6±0.6	**62.0±0.9**
Gemini 2.5 Pro	84.4±0.2	68.1±0.5	76.0±0.4	61.7±0.7
Gemini 3.1 Flash Lite	82.3±0.2	62.2±0.3	74.5±0.7	53.3±0.4
Gemini 3.5 Flash	**84.8±0.0**	**68.6±0.1**	78.3±0.3	61.0±0.3

to defeat surface cues, and any question that blind LLMs answered correctly at least 60% of the time was flagged for manual review and then dropped or rewritten. Each question is tagged by the modality of evidence it requires to answer correctly (*video*, *EHR*, and *cross-modal*). During evaluation we report the average scores across 5 runs, with each run using a different position for the ground truth answer. All reported results use a single fixed set of prompts and tools. The agent prompt development procedure is described in Appendix A.4. The report prompt, tool set, atomic action taxonomy, and temporal hierarchy were designed without reference to either benchmark. Example questions are in Appendix A.1.

Baselines. For the *text-only LLM* baseline, the agent has no access to any data and attempts to answer the question solely based on language priors. *Long-context VLMs* receive the question together with 475 or 950 frames sampled uniformly across the entire stay (Qwen3-VL 8B; Gemini 3.1 Flash Lite). *Socratic Models* (Zeng et al., 2023) baselines follow the protocol proposed in HourVideo (Chandrasegaran et al., 2024), where a VLM captions every minute of the stay and those captions are used by a language model to answer a question. Captions are temporally subsampled if their aggregate exceeds the LLM context length. We also show results for both long-context and Socratic baselines when including the stay's EHR in context (ranging from 130k-420k tokens). We note that these baselines use roughly three times more tokens than TRACER during question answering while achieving substantially lower accuracies. We include a more in depth discussion about baselines and compute budget in Appendix A.7. Baselines use their own fixed prompts, and additional implementation details for each baseline are included in Appendix A.7.

4 Experiments and Results

In order to evaluate TRACER as a harness, we perform experiments at three levels of granularity. First, we describe the experiments and results for minute-level perception evaluation in Section 4.1. Section 4.2 describes the experiments and results on our TRAJECTORY-144 long-form MCQ benchmark. Finally, we report our findings when applying TRACER across our entire 280 patient cohort in Section 4.3.

4.1 Minute-level Perception

Table 2 summarizes aggregate performance on our atomic action benchmark spanning 397 one-minute clips. Gemini 3.5 Flash performs best among the evaluated models, achieving 84.8% accuracy and a micro-F1 score of 68.6. Gemini 3.1 Flash Lite achieves 82.3% accuracy and a micro-F1

Table 3: **Harness ablation.** Accuracy (%) averaged over three agent models: Gemini 3.5 Flash, Gemini 3.1 Flash Lite, and Qwen3.5 9B. Single-toolset rows give the agent one tool family; cumulative rows add one family at a time up to the full harness, with gains shown in green. Best per column and block in bold, second best underlined. Additional results in Appendix A.12 and A.13.

	Video			**EHR**	**Cross-modal**	**Avg.**	
Agent tools	<5 min	5–30 min	>30 min	Record only	Video + EHR		
Single toolset performance							
Timeline only	27.5	43.3	**60.7**	30.3	25.0	<u>43.7</u>	
Reports only	<u>33.7</u>	**52.5**	<u>47.4</u>	<u>33.3</u>	<u>43.3</u>	42.1	
EHR only	32.9	31.7	41.2	**96.4**	**60.0**	**49.3**	
Video only	**39.8**	<u>50.8</u>	38.3	25.1	27.5	37.0	
Cumulative toolset performance							Δ
w/ timeline	27.5	43.3	60.7	30.3	25.0	43.7	0.0
+ reports	30.5	<u>55.4</u>	63.8	31.3	31.7	47.6	+3.9
+ EHR	<u>31.7</u>	51.6	<u>64.1</u>	**97.2**	**61.7**	<u>61.1</u>	+17.4
+ video (TRACER)	**39.2**	**57.1**	**69.2**	<u>96.9</u>	<u>56.7</u>	**65.3**	+21.6

score of 62.2 at less than one-tenth the cost, motivating its usage as the perception VLM when applying TRACER across our 280 patient cohort. We include more details about the cost-performance tradeoff in Appendix A.9. We show per-category performance of selected models in Figure 2(a).

4.2 Longitudinal Reasoning on TRAJECTORY-144

Comparison with baselines. We note the performance of Gemini 3.1 Flash Lite only reaches 45.0% using the Socratic models approach, only 49.3% when using raw video frames and EHR, and climbs to 61.0% when using TRACER. Baselines without the EHR score within 2.8 points of the text-only LLM. Although access to EHR raises the overall accuracy of baselines, they still score at most 83.8% on EHR-related questions, while TRACER scores 96.2%. With the same language model and the same data sources, TRACER's hierarchy and tool use outperform the long-context and Socratic baselines (+11.7 points over the strongest baseline; 95% CI [4.4, 18.9], paired bootstrap over questions).

Harness ablation. We ablate the harness across three different agent models under two settings (Table 3): single toolset and cumulative. In our single toolset ablation, we demonstrate that each toolset contributes most to performance for specific types of questions (Table 3, top). EHR tools significantly improve the performance on EHR-related questions, video tools allow for greater understanding of short moments in time spanning <5 minutes, reports allow for the agent to more accurately answer questions spanning 5 to 30 minutes in length, and the timeline tools are most helpful for questions that require reasoning across longer time horizons. Adding toolsets cumulatively (Table 3, bottom) yields sustained gains in average accuracy with each toolset added, and the full harness exceeds the best single toolset by 16 points. The best performance achieved by TRACER is using Gemini 3.5 Flash where it reaches up to 76.9% on TRAJECTORY-144. Additional per-agent results, paired tests, and per-stay breakdowns are in Appendices A.12 and A.13.

Agentic performance. To isolate the contributions of the agent LLM from the perception layer and temporal abstraction process, we conduct an experiment using Gemini 3.1 Flash Lite for video perception and temporal abstraction (minute-level record generation, hour reports, and all video-tool queries) while varying the agent LLM. Under this setting, open-weight models including Qwen3.5 9B and Gemma-4 31B match or outperform Gemini 3.1 Flash Lite as the agent LLM (Table 4). Together with the action detection results in Table 2, our findings indicate that the performance gap for evaluated open-weight models exists primarily from limitations in perception rather than agentic reasoning.

Table 4: **Comparison across agent models with fixed perception capabilities.** Accuracy (%) on TRAJECTORY-144 with Gemini 3.1 Flash Lite used for all video tool calls, timeline, and report construction. Best per column in bold, second best underlined.

	Video			EHR	Cross-modal	Avg.	Calls
Agent model	<5 min	5–30 min	>30 min	Record only	Video + EHR		
Proprietary							
Gemini 3.5 Flash	35.8	<u>57.5</u>	**80.3**	<u>98.5</u>	**77.5**	**70.7**	9.5
Gemini 3.1 Flash Lite	<u>36.4</u>	52.5	63.7	96.2	45.0	61.0	8.9
Open-weight							
gpt-oss-20b	32.1	41.2	62.5	93.1	40.0	57.4	3.7
Qwen3.5 9B	35.2	45.0	67.3	95.4	<u>50.0</u>	61.6	9.0
Gemma-4 31B	**40.0**	**61.2**	<u>73.7</u>	**100.0**	47.5	<u>67.9</u>	5.4

4.3 Cohort-Level Behavioral Signal

To determine whether TRACER can surface clinical signal at scale, we run its minute-level perception over the full cohort of 280 patients (roughly 60,000 camera-hours) and compare the frequency of each atomic action between patients with at least one positive delirium assessment during their stay ($n = 53$) and those with none ($n = 227$). We investigate only the 13 actions with sufficient positive samples in the perception benchmark, excluding other torso movements and clinician restraining. For each action we report the *delirium frequency ratio*, the per-minute frequency in the delirium group divided by that in the non-delirium group (see Appendix A.14 for more details). We visualize each evaluated atomic action's delirium frequency ratio in Figure 2(b). Seven actions are significantly less frequent among patients with documented delirium, and all but head turning remain significant after Holm correction on per-patient rates (Appendix A.14). The largest reductions are in purposeful, goal-directed actions: using non-medical objects ($0.41\times$), leg movements associated with daily living ($0.46\times$), object exchange with a clinician ($0.48\times$), and gesticulation ($0.60\times$). The fraction of observed minutes containing any of these four actions is lower in the delirium group (median 0.053 vs. 0.146 per patient, Mann–Whitney $p = 1.3 \times 10^{-8}$). Clinician–patient contact is the only atomic action that is more frequent for the delirium group ($1.20\times$), consistent with greater bedside attention. We include patient-level bootstrap intervals and an equal-patient-weight sensitivity analysis in Appendix A.14.

5 Related Work

Ambient intelligence in hospitals. Computer vision has been applied to inpatient monitoring through depth sensors and accelerometry for ICU mobility (Ma et al., 2017; Mehra et al., 2017), supervised detection of patient mobilization (Yeung et al., 2019), pervasive multimodal sensing (Davoudi et al., 2019), a supervised behavioral atlas of 40 ICU activities (Dai et al., 2025), and commercial real-time video analysis on general wards (Gabriel et al., 2025). These works exist within the broader research direction of ambient intelligence for healthcare (Haque et al., 2020; Yeung et al., 2018). In general, these systems train task-specific models on curated annotations using real patient data for a fixed set of behaviors, so the resulting models are fit to protected health information and are difficult to share publicly. In comparison with prior work, TRACER proposes a shareable harness for pretrained models and fuses the activity detections with the EHR to support question-answering over entire stays.

Long-video understanding. Video understanding benchmarks have recently extended from minutes (Mangalam et al., 2023) to an hour or more in length (Fu et al., 2024; Wu et al., 2024; Chandrasegaran et al., 2024). Existing long-form video understanding systems are generally designed for that regime and either compose frozen models through language (Zeng et al., 2023; Zhang et al., 2023; Wang et al., 2025), let an agent retrieve and re-inspect segments (Wang et al., 2024; Fan et al., 2025; Zhang et al., 2025), or maintain a multi-level caption memory with backtracking (Zuo et al., 2025). Week-long egocentric life-logs are the closest precedent to our work: EgoLife summarizes captions at hourly and daily levels and retrieves from day to hour to clip, with a fine-tuned captioner (Yang et al., 2025), while Ego-R1 trains an agent by reinforcement learning to navigate

a week/day/hour hierarchy (Tian et al., 2025). TRACER differs from existing works in both our setting and by fusing a second asynchronous modality (EHR) into our temporal hierarchy.

Clinical agents. LLM agents in medicine generally operate over text or structured data. Existing works focus on multi-agent question answering (Tang et al., 2024), code generation over electronic records (Shi et al., 2024), tool orchestration over single images (Fallahpour et al., 2025), and diagnostic dialogue (Tu et al., 2025). Recent work finds that LLM agents struggle to understand ICU stays from the EHR alone (Shen et al., 2026). TRACER extends upon prior work by examining the complementary information provided by days of continuous bedside video alongside EHR, rather than viewing each modality in isolation.

6 Discussion

Key findings. VLMs detect ICU atomic actions zero-shot with useful, if somewhat noisy, accuracy (68.6 F1). We demonstrate that these predictions, when combined with the EHR, can be used to construct a training-free LLM harness for greater understanding of patient stays. Specifically, TRACER enables LLM agents to more accurately answer questions over multi-day stays where existing approaches struggle (61.0% vs. at most 49.3% for the same backbone). When ablating our harness, we find that each toolset is best for different question archetypes. When run over the full 280-patient cohort, we show that the ICU atomic action detections provide signal for separating patients by delirium status. Purposeful actions such as using objects with hands and gesticulation are significantly less frequent among patients with documented delirium. Finally, we find that the gap between open-weight and proprietary models lies primarily in perception. When the perception layer is fixed, open-weight agents can match or even exceed the performance of proprietary models (Table 4). This has significant implications for on-premise deployment, which we find is currently limited more by the performance of the perception VLM than by the agent model.

Behavioral signal and delirium. The cohort-level associations shown in this work are consistent with the motoric phenomenology of delirium: patients with a documented positive screen show fewer purposeful, goal-directed actions. Hypoactive delirium, defined by reduced psychomotor activity, is both the most prevalent subtype in the ICU (Krewulak et al., 2018) and the one most often missed by intermittent screening. Furthermore, the sensitivity for its detection has been shown to degrade in routine practice (van Eijk et al., 2011). We emphasize that these are associations computed from model predictions on retrospective data; they are hypothesis-generating and motivate future work but do not by themselves establish predictive capabilities.

User Interface. Because ICUs already suffer from alarm fatigue (Sendelbach & Funk, 2013), TRACER is designed to give clinicians new capabilities for retrospective querying rather than serving as an engine for continuous alerting. The harness was built to run inside a viewer for bedside video data, which displays the detected atomic actions as a timeline underneath the video and lets clinicians ask questions in a chat panel. Clinicians can see the agent's tool calls and can click any cited timestamp to skip to the corresponding minute of video. We have demonstrated these capabilities to ICU nurses and physicians and used their feedback to guide the interface design.

Limitations. This is a single-center retrospective study; camera placement, population, and documentation practice may not transfer. TRAJECTORY-144 is deep rather than broad: 144 questions over two different ICU stays. Delirium labels come from intermittent bedside tests, and per-action detection error propagates into the detected delirium frequency ratios. Hourly reports are model-generated and may contain inaccuracies, hallucinations, or mistakes (see Appendix A.10). Continuous monitoring raises consent, staff-surveillance, and fairness questions (Haque et al., 2020; Martinez-Martin et al., 2021) that we did not explicitly discuss in this work but are prerequisites for deployment.

Reproducibility. The harness, tools, prompts, and taxonomy are released at `https://anonymous.4open.science/r/tracer-release-03B2`. Because we do not train any part of our harness on protected health data, the same hierarchy can be instantiated over any other temporally indexed video or health record collection. We include instructions for how to run our harness on sample videos and data in the linked repository.

AI USE STATEMENT

Large language model assistants (ChatGPT, Claude, and Claude Code) were used in preparing this work, in accordance with the ICLR AI Policy for Authors. *Writing:* AI assistants drafted and revised parts of the manuscript, edited text for readability and length, suggested structure for sections, converted the manuscript between formatting templates, formatted and verified bibliography entries, and checked consistency of terminology, numbers, and cross-references. *Code and experiments:* AI assistants helped implement the methods and baselines described in the paper, data cleaning and reformatting, debugging, and analysis code. The research questions, harness design, and atomic action taxonomy were developed by the authors and clinical collaborators, and all annotations were produced manually. TRAJECTORY-144 questions were written by annotators in consultation with clinicians from manually constructed event timelines; AI assistants were used to help draft and revise the wording of some questions and distractors, but every question was reviewed and verified by the authors. Every AI-assisted passage in the paper was reviewed and edited by the authors, who take full responsibility for the content.

REFERENCES

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

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. HourVideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

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

Wei Dai, Ehsan Adeli, Zelun Luo, Dev Dash, Shrinidhi Lakshmikanth, Zane Durante, Paul Tang, Amit Kaushal, Arnold Milstein, Li Fei-Fei, et al. Developing ICU clinical behavioral atlas using ambient intelligence and computer vision. *NEJM AI*, 2(2):AIoa2400590, 2025.

Anis Davoudi, Kumar Rohit Malhotra, Benjamin Shickel, Scott Siegel, Seth Williams, Matthew Ruppert, Emel Bihorac, Tezcan Ozrazgat-Baslanti, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. *Scientific Reports*, 9(1):8020, 2019.

John W Devlin, Yoanna Skrobik, Céline Gélinas, Dale M Needham, Arjen JC Slooter, Pratik P Pandharipande, Paula L Watson, Gerald L Weinhouse, Mark E Nunnally, Bram Rochwerg, et al. Clinical practice guidelines for the prevention and management of pain, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. *Critical care medicine*, 46(9): e825–e873, 2018.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical reasoning agent for chest x-ray. In *International Conference on Machine Learning*, 2025. arXiv:2502.02673.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2025.

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, Rongrong Ji, and Xing Sun. Video-MME: The first-ever comprehensive evaluation of multi-modal LLMs in video analysis. *arXiv preprint*, 2024.

Paolo Gabriel, Peter Rehani, Tyler Troy, Tim Wyatt, Michael Choma, and Narinder Singh. Continuous patient monitoring with AI: real-time analysis of video in hospital care settings. *Frontiers in Imaging*, 4:1547166, 2025.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6047–6056, 2018.

Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202, 2020.

Karla D Krewulak, Henry T Stelfox, Jeanna Parsons Leigh, E Wesley Ely, and Kirsten M Fiest. Incidence and prevalence of delirium subtypes in an adult ICU: a systematic review and meta-analysis. *Critical care medicine*, 46(12):2029–2035, 2018.

Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. Meta-Harness: End-to-end optimization of model harnesses. *arXiv preprint arXiv:2603.28052*, 2026.

Andy J Ma, Nishi Rawat, Austin Reiter, Christine Shrock, Andong Zhan, Alex Stone, Anahita Rabiee, Stephanie Griffin, Dale M Needham, and Suchi Saria. Measuring patient mobility in the ICU using a novel noninvasive sensor. *Critical Care Medicine*, 45(4):630–636, 2017.

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

Nicole Martinez-Martin, Zelun Luo, Amit Kaushal, Ehsan Adeli, Albert Haque, Sara S Kelly, Sarah Wieten, Mildred K Cho, David Magnus, Li Fei-Fei, Kevin Schulman, and Arnold Milstein. Ethical issues in using ambient intelligence in health-care settings. *The Lancet Digital Health*, 3(2): e115–e123, 2021.

Rishab Mehra, Gabriel M Bianconi, Serena Yeung, and Li Fei-Fei. Depth-based activity recognition in ICUs using convolutional and recurrent neural networks. In *Mach. Learn. Healthc. Conf*, pp. 1–9, 2017.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, February 2026.

Sue Sendelbach and Marjorie Funk. Alarm fatigue: a patient safety concern. *AACN Advanced Critical Care*, 24(4):378–386, 2013.

Chengzhi Shen, Weixiang Shen, Tobias Susetzky, Cherise Chen, Jun Li, Yuyuan Liu, Xuepeng Zhang, Zhenyu Gong, Daniel Rueckert, and Jiazhen Pan. RealICU: Do LLM agents understand long-context ICU data? a benchmark beyond behavior imitation. *arXiv preprint arXiv:2605.13542*, 2026.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-R1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 2025.

Maarten M van Eijk, Mark van den Boogaard, Rob J van Marum, Paul Benner, Piet Eikelenboom, Marina L Honing, Ben van der Hoven, Janneke Horn, Gerbrand J Izaks, Annemieke Kalf, et al. Routine use of the confusion assessment method for the intensive care unit: a multicenter study. *American Journal of Respiratory and Critical Care Medicine*, 184(3):340–344, 2011.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2405.19209.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2503.03803.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2405.15793.

Serena Yeung, N Lance Downing, Li Fei-Fei, and Arnold Milstein. Bedside computer vision—moving artificial intelligence from driver assistance to patient safety. *New England Journal of Medicine*, 378(14):1271–1273, 2018.

Serena Yeung, Francesca Rinaldo, Jeffrey Jopling, Bingbin Liu, Rishab Mehra, N Lance Downing, Michelle Guo, Gabriel M Bianconi, Alexandre Alahi, Julia Lee, Brandi Campbell, Kayla Deru, William Beninati, Li Fei-Fei, and Arnold Milstein. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *npj Digital Medicine*, 2(1):11, 2019.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.18079.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. VideoLucy: Deep memory backtracking for long video understanding. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2510.12422.

A APPENDIX

A.1 TRAJECTORY-144 EXAMPLE QUESTIONS

We include the following de-identified questions from TRAJECTORY-144. We show two questions per archetype. Calendar dates are replaced by the day of the stay and clock times by [time].

Video moment (<5 min) stay B

At [time] on day 3 of the stay, what clinical object was removed from the patient?
(A) An SCD sleeve
(B) An endotracheal tube
(C) The ventilator (extubation)
(D) A heel protector boot ✓
(E) A central line

Video moment (<5 min) stay B

From the observable video, around [time] on day 1 of the stay, what kind of bedside imaging was performed?
(A) An MRI
(B) A PET scan
(C) An ultrasound scan ✓
(D) A CT scan
(E) An X-ray

Video window (5–30 min) stay B

Over the 10 minutes from about [time] on day 2 of the stay, the patient was predominantly:
(A) Being handled by staff at the bedside (repositioning, linen or line adjustments), with only slight movements of her own
(B) Completely still, with no head, eye, or limb movement at all ✓
(C) Awake and talking with staff, gesturing with her hands
(D) Still except for occasional small hand or finger movements
(E) Repeatedly moving her arms and head on her own with no staff at the bedside

Video window (5–30 min) stay A

Between about [time] and [time] on day 2 of the stay, no staff were in the room. Which of these most accurately describes the patient's behavior?
(A) Repeatedly pulling at the tubing at her face until a nurse came in
(B) Talking with a visitor seated at the bedside
(C) Awake and using a smartphone while lying in bed ✓
(D) Sitting up on the edge of the bed
(E) Asleep with her eyes closed and her hands under the blanket

Long-form video (>30 min) stay B

Which of the following sequences correctly describes the order of events during the ICU recording?
(A) Patient taken out for tests and later returned → Ultrasound/scanning performed → Family members enter the room → SCD compression sleeves placed on the legs
(B) Ultrasound/scanning performed → SCD compression sleeves placed on the legs → Family members enter the room → Patient taken out for tests and later returned
(C) Ultrasound/scanning performed → Family members enter the room → Patient taken out for tests and later returned → SCD compression sleeves placed on the legs ✓
(D) SCD compression sleeves placed on the legs → Ultrasound/scanning performed → Family members enter the room → Patient taken out for tests and later returned
(E) Family members enter the room → Ultrasound/scanning performed → Patient taken out for tests and later returned → SCD compression sleeves placed on the legs

Long-form video (>30 min) stay B

How did the patient's level of voluntary movement and interaction before being taken out for tests compare to after returning to the room?

(A) The patient did not visibly interact before the tests but later spoke with clinicians after returning to the room.
(B) The patient appeared active and responsive both before and after the tests.
(C) The patient was mostly motionless before the tests but later sat up in a slanted position in the bed after returning to the room.
(D) The patient remained largely motionless both before and after the tests, with no clear voluntary movement observed. ✓
(E) The patient spoke with family members before being taken for tests and later spoke with clinicians after returning to the room.

EHR stay A

Besides sedatives and pain medication, the patient received a class of scheduled medications throughout the stay. Which class?

(A) Antibiotics
(B) Blood thinners (anticoagulants)
(C) Corticosteroids (steroids)
(D) Insulin
(E) Anti-seizure (antiepileptic) medications ✓

EHR stay A

The first CAM-ICU delirium assessment, on the first night of the stay, returned what result?

(A) Positive (delirium detected)
(B) Negative (no delirium detected)
(C) Unable to assess (UTA) ✓
(D) Negative on the initial screen, then re-assessed as positive later the same shift
(E) The assessment was attempted but the patient's RASS was too low (-4 or -5) to proceed

Cross-modal stay A

During the evening of day 2 of the stay, there was a stretch of about ten minutes with no staff in the room during which the patient used a smartphone while lying in bed. At about what time was that stretch, and which medication was administered within 45 minutes after it ended?

(A) [time window]; the norepinephrine infusion was stopped afterwards
(B) [time window]; no medication followed within 45 minutes
(C) [time window]; an albuterol–ipratropium nebulizer treatment followed ✓
(D) [time window]; a norepinephrine rate change followed
(E) [time window]; a chlorhexidine oral rinse followed

Cross-modal stay B

During the morning of day 2 of the stay, when did the patient have her period(s) of wakefulness?

(A) Multiple short awakenings starting around [time], each lasting only a few seconds
(B) Briefly around [time] and again around [time], with sedation in between ✓
(C) Around [time] only, with no other alert periods that morning
(D) A single stretch from about [time] through late morning
(E) Briefly around [time] only, then remaining sedated through the afternoon

A.2 Tool Interface and Agent Prompt

Tools. Table 5 lists every tool with its exact name, parameters (a question mark marks an optional parameter; `pat_id` defaults to the current patient and is omitted from the table except where it is the only parameter), and return value. We group tools by the families used in the ablations. Each tool configuration ablated in Table 3 consists of the scaffold tool set plus the families listed. Times are given in the hospital's local time (the zone name in tool names and prompt text is replaced by "local" for anonymization) and hours are counted from ICU admission.

Table 5: **The agent's tools.** Exact names, parameters, and return values of the eleven tools, grouped by family; the row labels of Table 3 name the families each configuration adds to the scaffold pair. A question mark marks an optional parameter.

Tool	Parameters	Returns
Scaffold (every configuration)		
`get_patient_metadata`	`pat_id`?	demographics, ICU admission time, observation window, encounter history
`convert_utc_local`	`dt_str`, `to_timezone`	the datetime converted between UTC and local time
Timeline (minute-level records)		
`get_atomic_actions`	`hour_index`	per-minute active action labels for one hour from admission
`query_actions`	`labels`?, `start_hour`?, `end_hour`?	minutes active per label over the range; per-hour counts; per-label per-hour minutes
`find_first_occurrence`	`labels`, `start_hour`?, `end_hour`?	first minute where any listed label is active (hour, minute, timestamp)
`find_moments_of_interest`	`target_labels`?, `start_pt`?, `end_pt`?, `min_rise_pct`?, `drug_window_min`?	rises and falls of a 15-minute rolling activity signal, with the medications given nearby
Report (hourly reports)		
`list_hourly_reports`	`pat_id`?	the (date, hour) pairs with a report
`get_hourly_report`	`date`, `hour`	the six-section report for that hour
EHR		
`query_medications`	`start_pt`, `end_pt`, `name_filter`?, `class_filter`?, `max_rows`?	medication administration events in the window: time, name, action, dose, route, class
`query_flowsheet`	`disp_name_filter`?, `start_pt`?, `end_pt`?, `max_rows`?	flowsheet rows (time, measurement, value, unit); without a filter, the measurement names available
Video		
`query_video_minute`	`minute_pt`, `prompt`	the perception VLM's answer to the prompt for each of the two camera views of that minute

Agent prompts. The box below is the agent system prompt for the TRACER harness. Reduced tool configurations receive the same prompt minus the lines of the tools they do not expose and minus any working rule that names an absent tool. The text-only baseline receives the short prompt shown afterwards. The same prompt is used for all results across every agent model in this work. During MCQ evaluation, the agent receives the question as the user message shown underneath the agent and blind-LLM prompts. Finally, we show the prompt for when a question exhausts the 25-call budget and a forced-choice message is shown to the agent. The instructions to always choose an option exist only for multiple-choice scoring, and in the clinical interface the agent may report that the evidence is insufficient.

Agent system prompt and tool descriptions, full harness (placeholders for patient metadata)

```
You are an analytic assistant embedded in a clinical video-viewer app for ICU patients.
  ↪ Answer the multiple-choice question about the patient currently on screen. Investigate
  ↪  with your tools BEFORE deciding; base your answer on what the tools actually return,
  ↪ not on priors.

Tools available to you:
- get_patient_metadata: Get demographics, ICU admission time, observation window length,
  ↪ and encounter history for a patient. Defaults to the patient currently on screen if
  ↪ pat_id is omitted.
- convert_utc_local: Convert between UTC and local time.
- list_hourly_reports: List all hourly markdown reports available for a patient. Each entry
  ↪  has a date (YYYY-MM-DD) and an hour (0-23). Defaults to the current patient if pat_id
  ↪  is omitted.
- get_hourly_report: Fetch the full markdown content of a single hourly report. Reports
  ↪ follow a fixed 6-section structure: Patient Movement Summary, Patient-Clinician
  ↪ Interaction, Significant Moments, Medication Context, Activity-Medication Alignment,
  ↪ Brief Summary. Defaults pat_id to the current patient.
- get_atomic_actions: Return minute-by-minute atomic action detections for one hour of the
  ↪ patient's stay. hour_index=0 is the first hour from ICU admission. Each minute lists
  ↪ currently-active action labels (from the atomic action taxonomy). Defaults pat_id to
  ↪ the current patient.
- query_actions: Aggregate atomic action counts across a range of hours. Returns: (a) total
  ↪  minutes each label was active over the window, (b) per-hour counts of minutes with
  ↪ any selected label active, and (c) per-label per-hour minutes active (dict[label] ->
  ↪ list over the hour range) -- use (c) to answer questions like 'which hour had the most
  ↪  X'. Filter by specific labels or omit to include all. start_hour/end_hour are zero-
  ↪ indexed from ICU start (end_hour is exclusive). Defaults pat_id to the current patient
  ↪ .
- find_first_occurrence: Find the earliest minute where any of the given action labels was
  ↪ active. Returns hour_index, minute (0-59), and UTC timestamp. Prefer this over reading
  ↪  hourly reports or scanning atomic actions when the user asks 'when did X first happen
  ↪ '. Optionally restrict to a range via start_hour/end_hour (zero-indexed from ICU start
  ↪ , end_hour exclusive). Defaults pat_id to the current patient.
- find_moments_of_interest: Detect 'moments of interest' -- rises and falls in a 15-minute
  ↪ rolling signal built from the atomic-action labels you select -- and cross-reference
  ↪ each detected blurp with medication events given nearby. The signal at minute t is the
  ↪  fraction of the previous 15 minutes in which any selected label fired. **You must
  ↪ choose target_labels to match the question.** Examples:
- query_medications: Return medication-administration-record (MAR) events for the current
  ↪ patient in a local-time window. Each event has taken_time_pt, med_name (display name),
  ↪  mar_action (Given / New Bag / Bolus / Stopped / etc.), dose (rate + unit or bolus +
  ↪ unit), route, and drug class (pharmacologic + therapeutic). Use this to answer
  ↪ questions like 'which meds were administered +/-45 min of X?', 'when was the last dose
  ↪  of oxycodone?', or 'what sedatives were running at 03:00?'. TIME: pass local-time
  ↪ strings 'YYYY-MM-DD HH:MM' (24-hour). Windows are inclusive on both ends. OPTIONAL
  ↪ FILTERS: name_filter (case-insensitive substring on the med name) and class_filter (
  ↪ substring on pharmacologic/therapeutic class, e.g. 'OPIOID', 'SEDATI', 'ANTIBIOTIC').
  ↪ Results are capped at max_rows (default 50).
- query_flowsheet: Query nursing-flowsheet measurements for the current patient: vitals (
  ↪ Heart Rate, Pulse, SpO2, Resp, Arterial Systolic BP, Mean Arterial Pressure, Arterial
  ↪ Diastolic BP, BP), sedation/agitation (Patient RASS Score, Anxiety Level, Pain Scale
  ↪ Type), and nursing assessments (Activity, ADLs, Communication, Ectopy, Cardiac Rhythm,
  ↪  etc.). Each row has recorded_time_pt, disp_name, value, and unit. DISCOVERY: call
  ↪ without disp_name_filter to see which measurement types exist for this patient (list
  ↪ of disp_name + count). QUERY: pass disp_name_filter (case-insensitive substring, e.g.
  ↪ 'RASS', 'Heart Rate', 'SpO2') to fetch the actual measurements. Optionally scope with
  ↪ start_pt / end_pt (local-time 'YYYY-MM-DD HH:MM'). Use this for questions like 'when
  ↪ was the patient most deeply sedated?' (RASS), 'what was the peak heart rate?', 'when
  ↪ did BP swings occur?', 'what was the anxiety/pain trend?'.
- query_video_minute: Look at the ICU surveillance video for ONE specific minute and answer
  ↪  a free-form question about what is visible -- patient movements/reactions, what a
  ↪ clinician is doing, devices/objects, body position, who is in the room.
  Treat its answer as EVIDENCE TO WEIGH, not ground truth: the video reader can miss brief
  ↪ events or misread motion, and the two camera views may disagree. When the answer
  ↪ matters, re-query (a different prompt or an adjacent minute) and cross-check against
  ↪ your other tools before committing.
  This tool is useful for zooming in on a single minute of interest. You can call it on the
  ↪  SAME minute several times with DIFFERENT prompts to probe different details (e.g.
  ↪ first ask what the clinician is doing, then whether the patient moved). It returns
  ↪ BOTH camera angles (left and right view) answered separately.
  TIME: everything here is local time. Pass the minute exactly as it appears in the
  ↪ question, 'YYYY-MM-DD HH:MM' (24-hour, local time); the tool finds the right clip for
  ↪ you and you never convert timezones. The clip is one minute long, so if an event might
  ↪  straddle a boundary, also query the adjacent minute.

If a tool returns an error or nothing useful, adjust and try another approach. Tools that
  ↪ take a patient id default to the current patient -- omit pat_id. You MUST select
```

```
  ↪ exactly one option as your single best guess, even if you are unsure or the data is
  ↪ incomplete. Never refuse and never say you cannot determine the answer.

Patient context (from get_patient_metadata):
- Video and activity data (hourly reports, atomic actions, query_video_minute) cover {start
  ↪ } to {end} local time ({n_hours} hours).
- hour_index counts whole hours from {start}: {hour_examples}. To convert a time to
  ↪ hour_index, subtract {start} and take the whole number of hours; write the subtraction
  ↪  down.
- EHR data (query_medications, query_flowsheet) covers the whole hospital admission, {
  ↪ ehr_start} to {ehr_end} local time, which extends beyond the video window.
Every date or time in a question refers to THIS stay in local time. Never guess a year or a
  ↪  date outside these windows.

Working rules (follow them strictly):
1. For "during the stay", "first", "last", "how many", "which day" questions about
  ↪ medications or flowsheet data, query the ENTIRE admission in ONE call (start_pt = {
  ↪ ehr_start}, end_pt = {ehr_end}); for activity data use start_hour 0 to end_hour {
  ↪ n_hours}. Do not step through the stay day by day.
2. Zero rows or zero events never prove absence. Before concluding "none", re-check the
  ↪ dates against the patient context, remove filters, widen the window, and try a shorter
  ↪  substring. If a result carries a "hint" field, follow it.
3. Medications: use name_filter ALONE with a short substring of the drug name (e.g. 'norepi
  ↪ ', 'oxycod', 'hepar'). Do not add class_filter unless the question is about a drug
  ↪ class, and never combine the two.
4. Flowsheet measurement names must be discovered, not guessed: call query_flowsheet with
  ↪ NO disp_name_filter first to see the exact names, then filter with a short substring
  ↪ taken from that list (e.g. 'CAM' rather than 'CAM-ICU', 'Temp' rather than '
  ↪ Temperature').
5. Outputs are capped. If n_events_returned < n_events_total, n_rows_returned <
  ↪ n_rows_total, or a "_truncated" note appears, narrow the filter or time window (or
  ↪ raise max_rows) before drawing any conclusion from a partial list.
6. Prefer totals the tools already computed (per_label_minutes_active, n_events_total,
  ↪ n_rows_total) over adding long arrays by hand. If you must count or add over a long
  ↪ list, do it in blocks of 10 and keep a running total.
7. For what the patient was doing over a several-minute window, first compute the
  ↪ hour_index, then use get_atomic_actions for that hour (minute-level labels) together
  ↪ with the hourly report; use query_video_minute to check the specific minute (and the
  ↪ adjacent minute) when the question is about a single moment.
8. Investigate with at least two different tools or queries before answering whenever the
  ↪ tools can bear on the question, and cross-check them against each other.
9. Keep the reasoning short and end with the line: Answer: <LETTER>
```

Agent system prompt, text-only configuration

```
You are answering a multiple-choice question about an ICU patient. You do NOT have access
  ↪ to the patient's video, activity data, or medical records, and you have no tools.
  ↪ Answer as best you can from general reasoning. You MUST select exactly one option as
  ↪ your single best guess, even if you are unsure or the data is incomplete. Never refuse
  ↪  and never say you cannot determine the answer.
```

User message for MCQ evaluation

```
QUESTION: {question}

OPTIONS:
{options, one per line as 'A. ...'}

You must choose exactly one option even if uncertain -- never refuse.
End your reply with a line in exactly this form:
Answer: <LETTER>
```

Forced-choice message (sent when the 25-call budget is reached)

```
You have used all available tool calls. Based on what you have found so far, commit to your
  ↪  single best-guess answer now. End with a line: Answer: <LETTER>
```

Video tool. `query_video_minute` sends the original one-minute clip of each camera view along with the same agent-written prompt, to the perception VLM (the agent's own model unless a row states otherwise) at 5 fps for Gemini models and 2 fps for open-weight models. A fixed

preamble precedes the agent's prompt on every call, and both views' answers are returned for the agent to reconcile. The tool's full description, which the agent reads through the tool definition, is the expanded `query_video_minute` entry in the system prompt box above; the box below is the fixed preamble the tool prepends to the agent's question before sending the clip.

Preamble prepended to the agent's prompt on every video tool call

```
You are an expert ICU video analyst. Patients here are often sedated, so subtle patient
  ↪ movement (a brief head turn, eye flutter, blink, cough, swallow, finger/limb twitch)
  ↪ is clinically relevant. Review the one-minute clip carefully and list ONLY movements
  ↪ you can actually see, each with an approximate timestamp (mm:ss). Do NOT infer or
  ↪ assume movements that are not clearly visible -- if the patient appears still, say so
  ↪ explicitly. Then answer the custom prompt below.
```

A.3 Temporal Hierarchy Construction Details

The perception VLM receives the 60 frames of a one-minute clip (1 fps) and a fixed prompt that requests a JSON object containing, for each of the 15 atomic actions, one of YES, NO, or UNABLE TO IDENTIFY, together with a brief free-text caption of the minute. The two camera views are merged by a logical OR over the 15 labels, with UNABLE TO IDENTIFY treated as negative, giving one binary label vector per patient-minute and the two view-specific captions are kept. Storing each minute as a 15-bit vector enables efficient timeline tools and quick scanning of multi-day stays. So, we collapse UNABLE TO IDENTIFY because it is relatively rare: Gemini 3.1 Flash Lite used it for 1 of 29,775 labels across five benchmark runs and in zero instances of 198,810 labels in a uniform sample of 13,254 minute records from the cohort run. On the benchmark an abstained prediction is excluded from the metrics rather than scored as negative. Among other perception VLMs, only Gemini 2.5 Flash and 2.5 Pro abstain often (3% of labels). Every other evaluated model abstains on under 0.5%. Sampling settings for every model are listed in Appendix A.8.

Perception prompt. The prompt below is sent with the 60 frames to every perception VLM, on the clip benchmark and on the cohort run (the clip identifier in the example is replaced by a placeholder). The user turn that follows it reads "Please find the video frames below:" and is followed by the frames.

Perception prompt (system turn)

```
MAIN INSTRUCTIONS:
Your objective is to analyze video frames extracted from a 1-minute Intensive Care Unit (
  ↪ ICU) hospital patient surveillance video recorded by a fixed, corner-mounted camera.
This is STRICTLY a FINE-GRAINED video understanding task.
You will receive 60 images sampled every 1 second, totaling 1 minute.
Your ONLY task is to decide, for each predefined category below, whether the corresponding
  ↪ patient action or patient-clinician interaction is present in the video. Then write a
  ↪ short 1-2 sentence summary of this video (STRICTLY keep this short) summary.
You must ONLY output a structured JSON object with categorical labels and Short summary.
KEY VISUAL TASKS (TAXONOMY)
PATIENT MOTION -- ARMS/HANDS
  1) Grabbing medical equipment
     - The patient is actively grasping or holding medical devices
       (e.g., ventilator mask, tubing, IV lines, monitor cables).
  2) Using or grabbing non-medical objects
     - The patient is using or holding non-medical items
       (e.g., cup, towel, tissue, phone, food, remote).
  3) Gesticulating
     - The patient is making expressive arm or hand movements while communicating
       (e.g., waving, pointing, hand gestures while speaking).
  4) Adjusting pillows or bedsheets
     - The patient pulls, pushes, or rearranges pillows, blankets, or sheets.
  5) Other arm movements
     - Any noticeable arm/hand motion such as fidgeting, lifting/dropping arms,
       scratching, or non-specific movements.
     - This category can co-occur with other arm/hand categories above.

PATIENT MOTION -- TORSO
  6) Patient changes body position in bed
     - The patient changes posture in bed
       (e.g., lying flat to leaning on one side, sitting up to lying down).
  7) Other torso movement
     - Other significant torso movements
       (e.g., quick or repeated body movements, jerking, agitation-like movements).

PATIENT MOTION -- HEAD/NECK
  8) Patient is looking toward other people in the room
     - The patient's gaze is clearly directed at another person.
  9) Patient turns or moves head/neck
     - Visible head or neck movement
       (e.g., turning left/right, nodding, tilting up/down).

PATIENT MOTION -- LEGS
  10) Leg movements related to mobility
      - The patient is making leg motions associated with activities of daily living
        (e.g., standing up, sitting down, walking, transferring to or from the bed).
  11) In-bed leg movement
      - The patient is making leg movements while they remain in bed
        (e.g., leg fidgeting, repositioning, leg movements under the covers).

PATIENT-CLINICIAN INTERACTIONS
```

```
  12) Clinician touches patient or uses tool/device on patient
      - A clinician or other person makes physical contact with the patient
        OR applies a tool or device directly to the patient's body.
  13) Patient communication
      - The patient appears to be speaking, mouthing words, or nodding in a way that
        responds to another person.
  14) Object exchange
      - An item is being handed from a person to the patient or from the patient
        to another person (e.g., cup, towel, tool, brush, phone).
  15) Patient is being physically restrained or stabilized
      - Another person is clearly holding or stabilizing part of the patient's body
        in a way that limits or controls movement.

DECISION RULES
- Each category is INDEPENDENT. Multiple categories can be YES simultaneously
  for the same video. For example, a patient may be grabbing medical equipment
  (cat 1) while also performing other arm movements (cat 5) at the same time.
  Evaluate each category on its own merits.
- Base ALL judgments only on what is clearly visible in the frames. - Do NOT infer pain,
  ↪ emotions, intentions, diagnoses, or clinical reasoning. - If visibility is poor or you
  ↪  are not confident, use "UNABLE TO IDENTIFY" for that label. - Every label MUST be
  ↪ exactly one of:
    - "YES"
    - "NO"
    - "UNABLE TO IDENTIFY"
- Do NOT add extra fields or narrative text.
OUTPUT RULES
- Return a single JSON object with fixed keys for the video ID and all labels. - Do NOT
  ↪ include explanations, timestamps, or free-text descriptions. - Use the exact key names
  ↪  and allowed values shown below.
EXAMPLE OUTPUT FORMAT

  {
    "video_id": "<date>/<time>_<room>_lt.mp4",
    "labels": {
      "patient_grabbing_medical_equipment": "NO",
      "patient_using_non_medical_objects": "NO",
      "patient_gesticulating": "NO",
      "patient_adjusting_bedding": "NO",
      "patient_other_arm_movements": "NO",
      "patient_position_change": "NO",
      "patient_other_torso_movement": "NO",
      "patient_looking_at_others": "NO",
      "patient_head_turning": "NO",
      "patient_leg_ADL_movement": "NO",
      "patient_leg_in_bed_movement": "NO",
      "clinician_patient_contact": "YES",
      "patient_communication": "NO",
      "clinician_patient_object_exchange": "NO",
      "patient_restrained": "NO"
    },
    "summary": "A clinician is present and makes physical contact with the patient, while
  ↪ the patient remains mostly still with no notable movements."
  }

You must ALWAYS follow this structure exactly.
```

Hourly report prompt. The report model receives the system turn below, then a user turn that fixes the six-section format, the hour's 60 minute-level records from both camera views, and the medication administration events of that hour.

Hourly report prompt (system turn)

```
You are an ICU behavioral summarization model.
You will receive up to 60 minutes of Task-1 video caption outputs (one JSON entry per
  ↪ minute)
from two camera views (left and right) of an ICU patient room,
plus a list of medication administration events relevant to this hour.

Your job is to generate a structured behavioral report for the
observation period based purely on visible evidence from both camera views.
You must never infer sedation level, intent, pain, emotion, or clinical state.

When both camera views are available for a given minute, reconcile them:
if either view detects an action, treat it as observed. Note discrepancies
only when they are informative (e.g., one view blocked by equipment).
```

```
When only one camera view is available, note which view and proceed.

TIMING REQUIREMENT (IMPORTANT):
  - All behavioral findings must cite timing.
  - Use minute timestamps (e.g., min 14, min 27) AND/OR continuous ranges (e.g., min 10-15)
  ↪ .
  - No event may be mentioned without a timing reference.

ALLOWED TIMING FORMATS:
  - "arm movement observed at min 12"
  - "clinician contact min 21-24"
  - "no movement detected min 0-18"

Your output must follow the 6-section template provided in the user prompt EXACTLY.
```

Hourly report prompt (user turn, before the hour's records)

```
Generate the behavioral report for this observation hour.

REQUIRED FORMAT:

Observation window: [START_TIME] to [END_TIME] (up to 60 minutes)

1. Patient Movement Summary
    - Overall movement level (high / moderate / low / none)
    - Most common movement types
    - MUST cite minute timestamps or ranges for all movement observed
    - Note any discrepancies between left and right camera views

2. Patient-Clinician Interaction Summary
    - Whether interaction occurred
    - MUST cite timings for clinician contact, communication, restraint, exchanges

3. Significant Moments of Interest
    - Sudden spikes in movement, first movement after stillness, notable behavior
    - MUST include timing (single minute or range)
    - If prolonged stillness >= 10 min, cite the stillness window

4. Medication Context
    - List medications active or administered during this hour with timestamps
    - Include drug name, dose, route, and action (Given/Rate Change/New Bag/etc.)
    - Note any continuous infusions that are running

5. Activity-Medication Alignment (observational only)
    - Timing-based associations ONLY (e.g., "movement at min 22-28 while dexmedetomidine
  ↪ infusing")
    - No clinical cause/effect interpretation ever

6. Brief Summary (<= 5 sentences)
    - Compact overall description of the hour
    - Note uncertainty if many labels = "UNABLE TO IDENTIFY"
    - Note if camera coverage was incomplete

Output MUST contain all 6 sections and MUST cite times when describing events.
```

A.4 TRAJECTORY-144 QUESTION CREATION AND FILTERING

We generated MCQs through a three-stage process. In the first stage, questions were created manually by human annotators with access to the underlying patient data. The question types were informed by discussions with clinicians, nurses, and domain experts in long-form video understanding. Each question was designed to assess one of three capabilities: (1) understanding changes in patient status over extended periods of time ($>$ 1 hour), (2) understanding fine-grained patient activities over shorter time spans (1 minute–1 hour), or (3) retrieving and reasoning over electronic health record (EHR) information to answer clinical queries accurately.

In the second stage, the question text was revised to increase their difficulty, and plausible but incorrect distractor answers were created. These distractors were designed to reduce the effectiveness of superficial linguistic cues and discourage models from exploiting language-based biases instead of reasoning over the underlying patient data.

Finally, we manually verified the correctness of each question and corresponding distractors. Any question that blind LLMs with no patient data answered correctly with at least 60% average accuracy was flagged for review, and flagged questions were dropped or reworked. Even after reworking, a number of questions remain guessable for the blind LLM baseline: with no patient data, 25 of the 144 questions (17%) are answered correctly at least 80% of the time by Gemini 3.1 Flash Lite. These questions account for part of the text-only baselines' accuracy above chance.

The final set of questions in TRAJECTORY-144 span temporal resolutions from concrete moments ("Around [time], what examination did the clinicians perform?") to whole days ("During the first 24 hours of the stay, did the patient leave her bed or her room?"), and also test reasoning over the record ("How many times was the patient formally assessed for delirium during the stay?").

Evidence types. For analysis every question is tagged by the evidence it requires. The 110 *video* questions are split by the span of video that resolves them: *moments* that a small number of clips can answer (under five minutes; 33 questions), *windows* of 5 to 30 minutes (16), and *long-form* spans of more than 30 minutes up to the whole stay (61). The remaining questions are answerable from the record alone (*EHR*; 26) or need both video and record to correctly answer (*cross-modal*; 8). The benchmark probes many questions over two long trajectories and does not test generalization across patients. We note that accuracy differs between the two stays, so Appendix A.12 reports each stay (denoted A and B) separately.

Temporal anchoring. Table 6 splits the questions by the time reference they give: 60 name a clock time, 34 an hour, part of a day, or a day, and 50 give none. The agent localizes events through a variety of mechanisms largely depending on the information present in the question: whether a time is provided in the question, whether it mentions a specific event present in the EHR, whether the question is able to be localized by an action query over the timeline, or by simply reading hourly reports over selected windows. The most specific temporal references generally belong to video-moment questions, where the difficulty generally lies in perception and not localization.

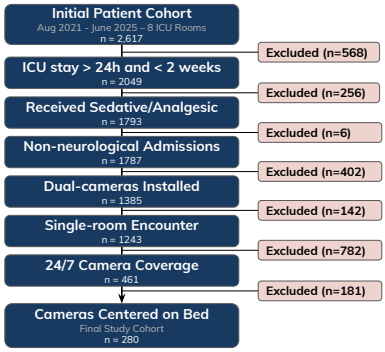
Agent prompt development. We refined the agent prompt using earlier draft versions of benchmark questions. None of the final 144 questions, in their evaluated form, was used for prompt development. All reported evaluations used a single fixed agent prompt.

Table 6: **TRAJECTORY-144 question breakdown by time reference.** Accuracy (%) for Gemini 3.1 Flash Lite using TRACER, from an independent run of the Table 1 configuration (achieves 61.1% overall).

Time reference in the question	Questions	Accuracy
Clock time, span under 5 minutes	33	30.9
Clock time, span 5–30 minutes	15	57.3
Clock time, span over 30 minutes	12	50.0
Hour	13	50.8
Part of day (e.g., morning, afternoon)	16	66.2
Day (e.g., November 23)	5	56.0
No explicit time, anchored on clinical record	22	96.4
No explicit time, anchored on video data	28	78.6

A.5 PATIENT COHORT

We visualize our clinical cohort and characteristics of our final 280 patient study cohort in Figure 3.



Characteristic	**Overall** ($N = 280$)
Age, years, median [IQR]	66.0 [58.1, 74.7]
Male, n (%)	183 (65.4)
BMI, kg/m^2, median [IQR]	25.8 [23.2, 29.9]
Admission type, n (%)	
Elective	140 (50.0)
Emergency	103 (36.8)
Urgent	30 (10.7)
Trauma Center	7 (2.5)
ICU LOS, days, median [IQR]	3.3 [2.0, 5.4]
Invasive mechanical ventilation, n (%)	157 (56.1)
In-hospital mortality, n (%)	13 (4.6)
CAM-ICU positive during stay, n (%)	53 (18.9)

Figure 3: **Patient cohort selection and characteristics. Left**, CONSORT-style flow diagram of patient cohort selection. Sequential exclusion criteria were applied to identify patients with sufficient ICU stay duration, sedative exposure, non-neurological admission, and adequate bilateral video coverage. **Right**, Characteristics of the final study cohort ($N = 280$).

A.5.1 PATIENT COHORT SELECTION

From the initial pool of 2,617 patients assessed for eligibility between August 2021 and June 2025, we applied sequential exclusion criteria to identify a cohort suitable for evaluating continuous video-based behavioral monitoring. Patients were required to have ICU stays exceeding 24 hours but less than two weeks (excluding 568), have received at least one sedative or analgesic medication (excluding 256), and have non-neurological admitting diagnoses (excluding 6). Technical requirements included admission to a room with two cameras installed (excluding 402), single-room encounters with no intra-ICU transfers (excluding 142), continuous bilateral video feed with 100% coverage over the entire stay (excluding 782), and confirmation that both cameras were centered on the patient bed area (excluding 181). See Figure 3 for a visualization.

A.5.2 ELECTRONIC HEALTH RECORDS

The harness reads four kinds of structured records, each through a specific component (tool signatures in Appendix A.2). (1) *Demographics and encounter metadata*: age, sex, race and ethnicity, height and weight, and per-encounter fields (admission type and source, hospital service, department and room, level of care, ICU days, discharge disposition), returned to the agent by the patient-metadata tool together with the ICU admission time and the video observation window. (2) *Admission, discharge, and transfer (ADT) events*: used to define ICU stays as contiguous periods at the "Critical Care" level of care, room assignments, and the observation window; the agent sees them only through the windows and room reported by the metadata tool. (3) *Medication administration records*: timestamped administrations with drug name, dose, route, action (Given, Rate Change, New Bag, Stopped, and similar), and pharmacologic and therapeutic class, exposed by the medication tool and supplied to the report model as the medication context of each hour. (4) *Nursing flowsheet measurements*: timestamped rows with a measurement name, value, and unit, covering vital signs (heart rate, blood pressure, oxygen saturation, respiratory rate), sedation and agitation scores (Richmond Agitation-Sedation Scale, anxiety, pain scales), Confusion Assessment Method for the ICU (CAM-ICU) delirium screens, and nursing assessments (activity, activities of daily living, communication, cardiac rhythm), exposed by the flowsheet tool, whose discovery mode lists the measurement names present for a patient. The EHR questions of TRAJECTORY-144 are answerable from (3) and (4); the cohort analysis of Section 4.3 takes its delirium labels from the CAM-ICU screens in (4) and the cohort characteristics of Table 15 from (1), (2), and (4).

A.6 TOKEN COST OF RAW VIDEO

Table 7 quantifies the coverage–detail trade-off described briefly in Section 2.1. For the one-minute clip the Gemini video tool sends the original 1920×1080 recording with 5 fps sampling. This results in 19,823 tokens per view, identical for both Gemini 3.1 Flash Lite and Gemini 3.5 Flash. We use 950 frames at 512×288 sampled uniformly across the stay for the long-context baseline, which uses 1,100 tokens per frame. For Qwen3-VL 8B, the same inputs were counted (patch size 16, 2×2 spatial merge, temporal patch size 2, frames used at native size): 144 tokens per individual frame and 8,640 tokens per one-minute clip. Under either tokenizer a single day of one camera exceeds the model's context by more than an order of magnitude. The 950-frame prompt still exists at the edge of a million-token window and, spread over a multi-day stay, samples one frame every several minutes. Thus, most events shorter than the sampling interval are likely to be missed. Table 1 shows the consequence, with every long-context baseline without EHR only able to land between 28.6% and 33.0%. By constructing the hierarchy as the stay is recorded and not using the entire hierarchy in context, each question requires a significantly smaller number of tokens.

Table 7: **Token cost of raw video surpasses context lengths.** Context windows: 1M tokens for the Gemini models, 256K for Qwen3-VL 8B. Gemini uses the original 1920×1080 clip at 5 fps. Qwen3-VL uses 512×288 resolution at 2 fps.

Input	Gemini 3 Family (tokens)	Qwen3-VL 8B (tokens)
One minute, one camera view	19,823	8,640
One minute, both views	39,646	17,280
One hour, both views	2.38M	1.04M
One day, both views	57.1M	24.9M
Median stay (3.3 days)	188M	82M
Longest stay in cohort (14 days)	799M	349M
TRAJECTORY-144 span (264 h)	628M	274M
One 512×288 frame	1,100	144
950 stills over a stay (long-context baseline)	1.05M	0.14M

A.7 BASELINE IMPLEMENTATIONS

All baselines answer the same number of questions (5 gold answer placements per question) as the harness and are scored identically. Two of the 144 questions have four options rather than five, so each configuration is scored on 718 trials. The confidence interval in Section 4.2 averages each question's placements, resamples the 144 questions with replacement (10,000 percentile-bootstrap resamples, the same questions for both systems); the McNemar tests in the appendix instead treat placements as independent trials and are therefore approximate.

Text-only. The text-only agent runs through the same agent loop with no tools and the short text-only system prompt of Appendix A.2, which states that it has no access to the patient's data. Its accuracy is intended to measure what language priors recover.

Socratic Models. Following the HourVideo protocol, a VLM captions every minute of the stay with the published HourVideo caption prompt, and a language model of the same family answers from the captions with the published HourVideo question prompt. This is a separate captioning pass from the harness's minute records: it uses a different prompt and produces free-text captions rather than action labels. Gemini 3.1 Flash Lite captions each clip at its default 1 fps and Qwen3-VL 8B is given 30 frames per clip. When the captions exceed the answering model's context they are subsampled at a fixed linear stride. For Gemini at 286 tokens per caption inside a million-token window, we sample every other minute for stay A (2,836 of 5,671 minutes) and every third for stay B (3,385 of 10,153). For the open-weight answerers of Table 9, tokens per caption range from 301 to 320, resulting in captions for 183 to 391 minutes per stay depending on the model's available context. Gemini baselines run at API defaults; the open-weight baselines use greedy decoding in bfloat16 (Appendix A.8).

Socratic baselines with clinical records and both cameras. The Socratic methods described in the paragraph above use one camera and no clinical records, whereas TRACER uses both cameras and can query the record. Thus, we add two Socratic variants with the same backbone (Gemini 3.1 Flash Lite as captioner and answerer) that close those gaps within a single million-token prompt (Table 8). Both prepend an *EHR block* to the captions: the patient metadata, every medication administration event, and every flowsheet variable of the stay. This block costs 131,000 tokens for stay A (670 flowsheet variables) and 420,000 for stay B (835 variables). The caption stride is re-derived so that block, captions, and prompts fit the model's context window. The first variant uses only the left-camera captions: every second minute survives for stay A, as before, and every sixth for stay B, resulting in prompts of 998,000 and 939,000 tokens. The second adds a right-camera caption for every minute, produced by the same prompt, model, and frame rate (15,824 clips), and keeps both captions of each surviving minute. This results in every fourth minute caption used for stay A and every eleventh for stay B and gives prompts of 689,000 and 706,000 tokens. Both prompts are several times the harness's per-question budget. This approach improves the performance on EHR questions from 16.2 to 76.2 and the average from 33.7 to 45.0 (McNemar over the 718 paired trials, $p < 0.001$), with video performance unchanged. Adding the second camera does not help further at a fixed budget (42.7) due to the larger temporal stride. Performance falls on long-form video questions (43.3 to 35.9) and moment and cross-modal questions move within noise. The harness with the same model reaches 61.0.

Long-context VLM. The model receives the question together with 475 or 950 frames under a fixed frames-only prompt. Frames are spaced uniformly over the minute clips of the stay, taking the midpoint of each selected clip, and are extracted from the left camera (the right camera is used only for minutes where the left clip is missing) and scaled so that the longer side is 512 pixels, giving 512×288. Frames are given in chronological order without timestamps. On Gemini 3.1 Flash Lite the 950-frame prompt uses 1,045,000 tokens (1,100 per frame), hitting the limit of its million-token window; 475 is half of that, and on Qwen3-VL 8B the same frames cost 144 tokens each. Over the two benchmark stays, 950 frames correspond to one frame every 6.0 and 10.7 minutes, so events shorter than that interval are likely to be missed by this baseline.

Long-context baseline with the clinical record. The same EHR block described above is added to the long-context baseline and we halve the frame count, resulting in 475 frames at 512×288

resolution from the left camera. The frames follow the EHR block in a prompt spanning 654,000 tokens for stay A and 943,000 for stay B. This is the strongest baseline and scores 49.3 (Table 8). Adding the EHR to the 475 frame baseline lifts long-form video performance from 36.6 to 48.9 as well as EHR performance from 17.7 to 83.8.

Table 8: **Baselines with clinical records using Gemini 3.1 Flash Lite.** Accuracy (%) averaged over gold placements. "Coverage" is the number of captioned minutes or frames that survive subsampling given the fixed 1M context for stay A / stay B; "Tokens" is the prompt length per trial. TRACER is shown on the last row for comparison.

Inputs	Coverage	Tokens	<5 min	5–30 min	>30 min	EHR	Cross	All
Left camera captions	2.8k / 3.4k	1.0M	34.5	32.5	39.2	16.2	47.5	33.7
Left camera captions + EHR	2.8k / 1.7k	1.0M / 0.9M	30.3	32.5	43.3	76.2	42.5	45.0
Both cameras' captions + EHR	1.4k / 0.9k	0.7M / 0.7M	32.7	31.2	35.9	76.2	50.0	42.7
950 frames	950 / 950	1.1M	35.8	32.5	38.0	14.6	25.0	31.9
475 frames	475 / 475	0.5M	26.7	35.0	36.6	17.7	22.5	29.9
475 frames + EHR	475 / 475	0.7M / 1.0M	34.5	30.0	48.9	83.8	40.0	49.3
TRACER harness, all tools (Table 1 run)	–	–	36.4	52.5	63.7	96.2	45.0	61.0

Open-weight question answering for Socratic baselines. Table 9 runs the Socratic protocol with the Gemini 3.1 Flash Lite captions for video perception and an open-weight model as the answerer. For each row, we show the performance of TRACER using the same model as the agent LLM. Every open-weight answerer lands within about ten points of chance when it can only read the captions, and gains 33 to 39 points when using the harness.

Table 9: **Socratic baselines with Gemini 3.1 Flash Lite captions, by open-weight answering model.** Accuracy (%) averaged over gold placements; "minutes" is the number of captioned minutes that survive subsampling for stay A / stay B; the last column is the same model as the TRACER agent with Gemini 3.1 Flash Lite serving the video tool. Served context length shown in parentheses.

Answering model	Minutes	<5 min	5–30 min	>30 min	EHR	Cross	All	TRACER
Gemini 3.1 FL (1M)	2.8k / 3.4k	34.5	32.5	39.2	16.2	47.5	33.7	61.0
gpt-oss-20b (131K)	379 / 391	33.3	26.3	22.7	14.6	20.0	23.9	57.4
Gemma-4 31B (131K)	379 / 377	33.9	25.0	31.4	23.1	22.5	29.3	67.9
Qwen3.5 9B (65K)	183 / 185	30.3	28.8	30.8	18.5	10.0	27.1	61.6

A.8 Model Instantiation and Computational Cost

All three roles are frozen, prompted models; their sampling settings are listed at the end of this section. For temporal abstraction we use Gemini 3.1 Flash Lite as both the perception and the report model. Although it is not the most accurate VLM we evaluate, it offered the best accuracy–cost trade-off for running over roughly 60,000 camera-hours. Hourly reports were generated, with the same model, only for the two TRAJECTORY-144 stays. To measure how perception quality varies across models we evaluate nine VLMs on the perception benchmark under the same prompt: the open-weight Qwen3-VL 8B (Bai et al., 2025), Qwen3.5 4B and 9B (Qwen Team, 2026), and Gemma-4 E4B (Team et al., 2026), and five Gemini models spanning three generations (Gemini Team, 2023; Comanici et al., 2025); our data use agreement permits processing protected health information among proprietary models only within the Gemini family. For the agent model we report Gemini 3.5 Flash as the best performing model, and generally use Gemini 3.1 Flash Lite as the reference configuration for most baseline comparisons.

Computational cost. Almost all visual computation happens in the background: constructing the temporal hierarchy costs two VLM calls per minute of stay and one LLM report call per hour. This cost is independent of how many questions are later asked. Answering a question is cheap by comparison. For Gemini 3.5 Flash, we find that the agent uses an average of 160,000 tokens per question over 9.8 turns. On average, each question uses 3.3 video calls adding approximately 132,000 tokens, resulting in roughly 292,000 per question. The distribution is heavy-tailed because a few trials reach the 25-call cap with long transcripts.

Sampling and serving settings. Gemini models are called with the API defaults for temperature, top-p, top-k, output length, and thinking in every role (perception, reports, agent, and baselines). Open-weight models are served with vLLM with thinking disabled. For the atomic action benchmark we use temperature 0.6, top-p 0.95, and top-k 20. For the long-context and Socratic baselines we use greedy decoding. Every agent runs under the same loop: the agent prompt, at most 25 tool calls, a forced answer when the budget is exhausted. The video tool sends the original one-minute clip with 5 fps sampling for Gemini models and 512×288 at 2 fps for open-weight models. The cohort perception run and the hourly reports use Gemini 3.1 Flash Lite at API defaults with the prompts of Appendix A.3.

A.9 Perception Benchmark Details

Accuracy-cost tradeoff. Figure 4 plots micro-F1 on the 397-clip benchmark against the cost of running each perception VLM over video. For the Gemini models, cost is computed from all recorded prompt and response tokens using the prices at the time of submission (September 2026). For the open-weight models, cost is measured wall-clock per clip on one AMD MI250X graphics compute die. Gemini 3.1 Flash Lite was selected to serve as the perception VLM for many of our experiments due to its strong and cheap performance at the Pareto frontier.

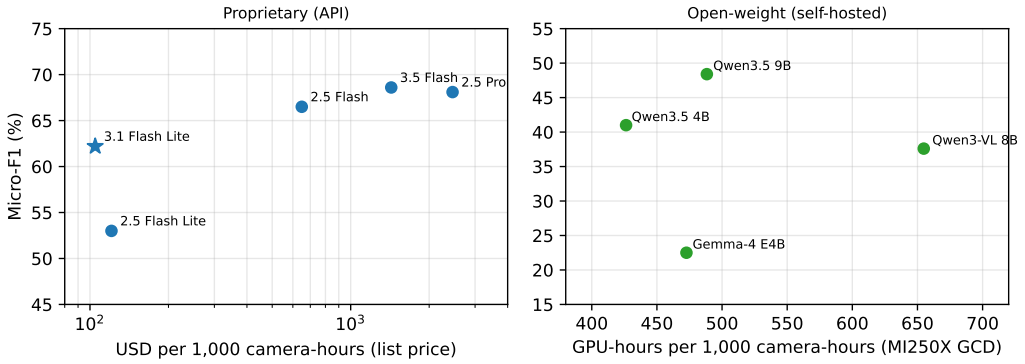


Figure 4: **Perception accuracy versus cost.** Left: Gemini models, list-price cost per 1,000 camera-hours of one-minute clips (log scale); the star marks the model used for the cohort run. Right: open-weight models, device-hours per 1,000 camera-hours on one MI250X graphics compute die.

Prevalence. In our 397-clip benchmark, category prevalence is imbalanced: the two most common actions (other arm movements, head turning) are present in $\sim$70% of the 397 clips, while the rarest appear in fewer than 5%. We note that the relative prevalence of atomic actions in the actual ICU video data is significantly lower for all categories due to our selection of clips with differing levels of patient activity, acuity, and time of day throughout the ICU stay.

Annotation reliability. In order to test annotation reliability, one author independently re-annotated 20 randomly sampled benchmark clips annotated by others, blinded to the original labels, model predictions, and clip metadata, using the finalized rubric. Agreement with the final verified benchmark labels, pooled across all 15 categories, was 92.7%, with $\kappa = 0.81$ (95% CI [0.74, 0.87], clip-level bootstrap; $N = 300$ paired decisions).

A.10 HOURLY REPORT FAITHFULNESS

Because the hourly reports are generated by a language model and then consumed as evidence by another, we audited ten reports, five per benchmark stay. Every claim that cites a minute, an action, a medication, or a temporal or causal relationship was extracted and checked against the minute-level records and the medication administration record. From our sample, we found 153 total claims (Table 10). Approximately 84% of these claims were clearly supported by the minute records and EHR. Most of the remainder (8%) were timing claims whose cited minutes only partly matched the minute records; six (4%) were statements that no medication was given during an hour in which the record shows one; the rest (4%) were overall movement ratings or absence claims that were contradicted by the minute records (e.g., describing an activity period as "low activity" despite many atomic action detections). Support was highest in the narrative summary and significant-moments sections and lowest in the movement and activity–medication sections. Every claim that named a medication matched the record on drug, dose, route, and minute. The dominant error was *false absence*: three reports stated that no medication was administered during an hour in which the record shows one, and each such statement propagated into the activity–medication section. We found no invented causal language in the sample. We note the limitations of our approach: a claim-level audit cannot detect omission. A larger audit covering per-hour completeness across a larger number of patients is necessary before deployment.

Table 10: **Claim-level audit of ten hourly reports.** Counts are verifiable claims. Partial timing match: under 80% of cited minutes show the action. Carry over: a false medication absence repeated in the alignment section.

	Report section							
	Movement	Interaction	Moments	Medication	Alignment	Summary	All	%
Verifiable claims	41	36	28	16	12	20	153	
Supported	31	32	25	13	9	19	129	84.3
Errors	10	4	3	3	3	1	24	15.7
Only partial timing match	7	1	3	0	0	1	12	7.8
Movement summary misleading	3	0	0	0	0	0	3	2.0
Incorrect action absence	0	3	0	0	0	0	3	2.0
Incorrect medication absence	0	0	0	3	0	0	3	2.0
Carry over of incorrect information	0	0	0	0	3	0	3	2.0
Support rate (%)	76	89	89	81	75	95	84	

A.11 How does TRACER Perform Across Different Agent Models?

Table 11 runs the full TRACER harness across 12 agent configurations. We find that Qwen3.5 9B's accuracy increases from 57.9 to 61.6 when using Gemini 3.1 Flash Lite for video perception. Gemma-4 31B is the strongest open-weight agent at 67.9, saturating the EHR questions and reaching 73.7 for long-form video questions. Gemini 3.5 Flash with Gemini 3.1 Flash Lite serving its video tool scores 70.7 against its original 76.9 (paired McNemar on 718 shared trials, 103 versus 59 discordant, $\chi^2 = 11.4$, $p < 0.001$), with the loss confined to video moments (54.5 to 35.8) and windows (75.0 to 57.5) while long-form and EHR questions are unchanged. This weaker perceiver also makes the agent re-query video more, hitting the 25-call budget on 10.7% of trials against none with its own perception.

Table 11: **TRACER applied across twelve agent configurations.** Full TRACER with the agent model as the only variable; the background records and reports are fixed. Tool calls are means per question. †Video tool served by Gemini 3.1 Flash Lite rather than the agent's own model.

Agent model	All	<5 min	5–30 min	>30 min	EHR	Cross	Avg. Calls
Gemini 2.5 Flash Lite	49.2	32.7	35.0	53.9	69.2	45.0	3.7
Gemini 2.5 Flash	60.3	33.9	56.2	64.0	90.8	50.0	4.7
Gemini 3 Flash	70.8	44.8	66.3	77.0	91.5	72.5	12.3
Gemini 3.1 Flash Lite	61.0	36.4	52.5	63.7	96.2	45.0	8.9
Gemini 3.5 Flash Lite	61.1	40.0	52.5	59.8	99.2	52.5	6.8
Gemini 3.5 Flash	**76.9**	**54.5**	**75.0**	**80.7**	99.2	72.5	9.0
Qwen3.5 9B	57.9	26.7	43.8	63.3	95.4	52.5	10.4
Qwen3-VL 8B	42.1	19.4	32.5	45.7	68.5	42.5	4.7
gpt-oss-20b†	57.4	32.1	41.2	62.5	93.1	40.0	3.7
Qwen3.5 9B†	61.6	35.2	45.0	67.3	95.4	50.0	9.0
Gemma-4 31B†	67.9	40.0	61.2	73.7	**100.0**	47.5	5.4
Gemini 3.5 Flash†	70.7	35.8	57.5	80.3	98.5	**77.5**	9.5

A.12 PER-MODEL CUMULATIVE ABLATION

Table 12 gives the cumulative ablation of Table 3 across many agent models separately.

Table 12: **Per-model cumulative ablation, by question type and by stay.** Cumulative ablation: each row adds one tool family to the row above (Table 5). Paired McNemar tests on identical trials for Gemini 3.5 Flash show that every step is significant ($p < 0.001$) except adding reports ($+3.5$, $\chi^2 = 3.6$, $p = 0.06$).

	Video			**EHR**	**Cross-modal**	**Stay**		**All**
Agent tools	<5 min	5–30 min	>30 min	Record only	Video + EHR	A	B	
Gemini 3.5 Flash								
Text-only	37.0	32.5	51.4	21.5	27.5	35.9	41.6	39.3
+ timeline	37.6	57.5	67.5	42.3	20.0	62.0	45.6	52.4
+ reports	40.6	72.5	70.5	38.5	30.0	70.2	45.9	55.8
+ EHR	39.4	76.2	71.1	100.0	75.0	84.7	59.5	69.9
+ video (TRACER)	**54.5**	75.0	**80.7**	99.2	72.5	**88.5**	**68.9**	**76.9**
Gemini 3.1 Flash Lite								
Text-only	24.2	18.8	40.2	26.9	25.0	26.8	33.8	30.9
+ timeline	23.0	32.5	55.3	26.2	22.5	50.5	29.8	38.3
+ reports	24.8	48.8	62.9	31.5	37.5	57.9	36.9	45.5
+ EHR	23.6	36.2	58.6	96.2	47.5	71.2	42.5	54.3
+ video (TRACER)	36.4	52.5	63.7	96.2	45.0	75.8	50.8	61.0
Qwen3.5 9B								
Text-only	31.5	27.5	43.9	20.0	40.0	35.3	34.4	34.7
+ timeline	21.8	40.0	59.3	22.3	32.5	52.1	32.2	40.4
+ reports	26.1	45.0	58.0	23.8	27.5	51.9	34.0	41.4
+ EHR	32.1	42.5	62.5	95.4	62.5	74.8	48.4	59.2
+ video (TRACER)	26.7	43.8	63.3	95.4	52.5	74.6	46.4	57.9
Qwen3-VL 8B								
Text-only	29.7	28.7	42.6	25.4	32.5	34.9	34.1	34.4
+ timeline	21.2	27.5	39.9	22.3	35.0	39.3	24.9	30.8
+ reports	19.4	48.8	46.7	18.5	22.5	39.5	30.6	34.2
+ EHR	17.0	33.8	43.5	70.0	25.0	53.8	30.6	40.1
+ video (TRACER)	19.4	32.5	45.7	68.5	42.5	55.7	32.7	42.1
gpt-oss-20b (video tool served by Gemini 3.1 Flash Lite)								
Text-only	22.4	17.5	36.5	22.3	42.5	27.5	29.9	28.9
+ timeline	22.4	36.2	56.5	16.2	35.0	47.4	31.4	38.0
+ reports	25.5	38.8	62.1	27.7	30.0	55.8	34.3	43.1
+ EHR	26.7	40.0	59.3	90.8	52.5	74.8	41.2	55.0
+ video (TRACER)	32.1	41.2	62.5	93.1	40.0	73.4	46.4	57.4

A.13 SINGLE TOOL FAMILY ABLATIONS

Table 13 gives the agent a single tool family on top of the scaffold tools, following the same protocol as Table 3.

Table 13: **Single tool families on TRAJECTORY-144.** Each row gives the agent one tool family only. Accuracy (%) averaged over gold placements; category columns are question means.

Agent tools	All	<5 min	5–30 min	>30 min	EHR	Cross
Gemini 3.5 Flash						
Reports only	51.0	44.2	72.5	55.4	36.9	47.5
EHR only	59.6	49.1	47.5	49.5	100.0	72.5
Video only	46.4	54.5	76.2	43.9	30.0	27.5
All tools	76.9	54.5	75.0	80.7	99.2	72.5
Gemini 3.1 Flash Lite						
Reports only	37.5	27.9	38.8	44.5	31.5	40.0
EHR only	45.3	26.1	20.0	40.7	95.4	47.5
Video only	34.5	38.2	45.0	36.8	22.3	20.0
All tools	61.0	36.4	52.5	63.7	96.2	45.0
Qwen3.5 9B						
Reports only	37.7	29.1	46.2	42.2	31.5	42.5
EHR only	43.0	23.6	27.5	33.5	93.8	60.0
Video only	30.2	26.7	31.2	34.3	23.1	35.0
All tools	57.9	26.7	43.8	63.3	95.4	52.5
gpt-oss-20b (video tool served by Gemini 3.1 Flash Lite)						
Reports only	32.8	26.7	43.8	33.9	30.0	37.5
EHR only	37.3	26.1	27.5	23.4	93.1	27.5
Video only	28.4	28.5	43.8	32.3	14.6	12.5
All tools	57.4	32.1	41.2	62.5	93.1	40.0

A.14 Delirium and Atomic Action Frequency Analysis

We define the *delirium frequency ratio* as being:

$$\mathrm{FR}_a = \frac{f_a^{\mathrm{del}}}{f_a^{\text{non-del}}}, \qquad f_a^{(g)} = \frac{\sum_{p \in \mathcal{P}_g} m_a^{(p)}}{\sum_{p \in \mathcal{P}_g} M^{(p)}}, \tag{1}$$

where $\mathcal{P}_g$ denotes the set of patients in group $g \in \{\mathrm{del}, \text{non-del}\}$, $m_a^{(p)}$ is the number of patient-minutes in which action a is detected for patient p, and $M^{(p)}$ is the total number of observed minutes for that patient; group frequencies are thus pooled over all observed minutes within each group (minute-weighted).

The delirium frequency ratios reported in Figure 2(b) and Eq. 1 are cohort-level associations, and several features of their construction warrant elaboration.

Group composition. Patients who screen positive for delirium may differ from those who do not along dimensions that independently suppress observable movement, including sedation depth, exposure to mechanical ventilation, and length of stay. Table 15 summarizes these characteristics by group. Ventilation exposure is similar between groups (invasive ventilation 58.5% vs. 55.5%), as is documented sedation level (per-patient median RASS 0.00 in both groups). The delirium group has a higher proportion of assessments indicating deep sedation (RASS ≤ -3: 5.0% vs. 2.0%) and a longer ICU stay (median 4.62 vs. 3.09 days), and is correspondingly observed for more minutes.

Weighting. Equation 1 pools numerator and denominator counts across all patients in a group, so each patient contributes in proportion to the number of minutes observed. Patients with longer stays therefore exert greater influence on group frequencies, and length of stay is itself associated with delirium. As a sensitivity analysis, we recomputed a different ratio that weights every patient equally—that is, by first computing a per-patient frequency $m_a^{(p)}/M^{(p)}$ and then averaging across patients within each group (Table 16). The three largest reductions are essentially unchanged under equal weighting: using non-medical objects ($0.41\times \rightarrow 0.39\times$), leg ADL movement ($0.46\times \rightarrow 0.47\times$), and object exchange ($0.48\times \rightarrow 0.47\times$). For this analysis, several ratios closer to unity became more frequent for delirious patients—adjusting bedding ($0.85\times \rightarrow 1.07\times$), position change ($0.85\times \rightarrow 1.06\times$), and in-bed leg movement ($1.01\times \rightarrow 1.30\times$). However, all of these show minimal differences between groups.

Uncertainty. To quantify sampling variability, we computed 95% confidence intervals for each $\log_2 \mathrm{FR}_a$ by resampling patients with replacement within each group (10,000 bootstrap replicates, patient-level resampling). Intervals are reported in Table 16. Of the 13 evaluated actions, seven have intervals excluding zero under *both* weightings and are reduced in the delirium group: using non-medical objects, leg ADL movement, object exchange, gesticulating, other arm movements, grabbing medical equipment, and head turning. One action, clinician–patient contact, is elevated in the delirium group under both the minute-weighted and patient-weighted estimators ($1.20\times$ and $1.18\times$). The remaining five—adjusting bedding, position change, looking at others, in-bed leg movement, and communication—are not distinguishable from no difference under either weighting, and the corresponding bars in Figure 2(b) should not be interpreted as evidence of a group difference. Ratios for infrequent actions and for categories with lower detection F1 carry correspondingly wider intervals and warrant greater caution.

Significance tests. The bootstrap intervals above cannot resolve small p-values, so Table 14 reports significance tests on the same cohort. For each action we compare per-patient rates (minutes with the action over observed minutes) between the 53 patients with a positive screen and the 227 without. We use a two-sided Mann–Whitney U test (tie-corrected asymptotic p, since rates are heavily tied at zero), adjusted across the 13 evaluated actions by Holm and by Benjamini–Hochberg. We test each $\log_2 \mathrm{FR}_a$ by permuting the group labels 2,000,000 times (two-sided, $p = (n_{\geq} + 1)/(B + 1)$, floor 5×10^{-7}). We define *purposeful activity* as the per-patient fraction of observed minutes containing any of the following four purposeful actions: using non-medical objects, object exchange, leg ADL movement, gesticulating. We find a median of 0.053 (IQR 0.024–0.118) in the delirium group against 0.146 (0.093–0.217) in the non-delirium group, a pooled rate of 0.080 against 0.161 ($\log_2 \mathrm{FR} = -1.01$), with $U = 2{,}998$, $p = 1.3 \times 10^{-8}$ by Mann–Whitney and $p = 5 \times 10^{-7}$ by permutation on $\log_2 \mathrm{FR}$. The three purposeful actions with the largest reductions remain significant at $p < 10^{-4}$ after Holm correction; gesticulating does not ($p = 1.5 \times 10^{-3}$ adjusted). Two actions

in the bootstrap-significant set are weaker under these tests: head turning is not significant on per-patient rates ($p = 0.094$ unadjusted), and grabbing medical equipment survives Holm correction only narrowly ($p = 0.034$). Clinician–patient contact, the one action more frequent in the delirium group, is robust ($p = 1.4 \times 10^{-5}$, Holm 1.4×10^{-4}).

Label timing. Patients are assigned to the delirium group if any CAM-ICU assessment during the stay is positive, whereas behavior frequencies are pooled over all observed minutes of the stay. The analysis therefore contrasts patients who experience delirium at some point against those with no documented positive screen, rather than contrasting delirious against non-delirious periods within a stay. Analyses aligned to assessment times—for example, restricting to windows adjacent to a positive screen, or comparing CAM-positive against CAM-negative days within the same patient—would more directly probe the motoric phenomenology of delirium and are a direction for future work.

Table 14: **Significance tests for delirium frequency differences.** Per-patient rates (median) by group with two-sided Mann–Whitney U p and Holm and Benjamini–Hochberg adjustment over the 13 evaluated actions. Permutation p for the minute-weighted and equal-weighted $\log_2$ frequency ratio over 2,000,000 label permutations (floor 5×10^{-7}).

Action	Med. del.	Med. non-del.	MWU p	Holm	BH	Perm. p (mw / ew)
Using non-medical objects	0.016	0.080	1.8×10^{-10}	2.3×10^{-9}	2.3×10^{-9}	5×10^{-7} / 5×10^{-7}
Object exchange	0.004	0.010	8.9×10^{-10}	1.1×10^{-8}	5.8×10^{-9}	5×10^{-7} / 5×10^{-7}
Leg ADL movement	0.002	0.007	7.3×10^{-7}	8.0×10^{-6}	3.2×10^{-6}	4×10^{-6} / 5×10^{-7}
Clinician–patient contact	0.423	0.349	1.4×10^{-5}	1.4×10^{-4}	4.7×10^{-5}	1.1×10^{-3} / 1.0×10^{-4}
Other arm movements	0.165	0.243	2.9×10^{-5}	2.6×10^{-4}	7.4×10^{-5}	6.5×10^{-6} / 2×10^{-6}
Gesticulating	0.030	0.053	1.9×10^{-4}	1.5×10^{-3}	4.1×10^{-4}	1.4×10^{-4} / 3.5×10^{-4}
Grabbing medical equipment	0.081	0.102	4.8×10^{-3}	0.034	8.9×10^{-3}	1.1×10^{-3} / 4.1×10^{-3}
Head turning	0.363	0.399	0.094	0.56	0.15	4.3×10^{-3} / 8.6×10^{-3}
Position change	0.027	0.029	0.11	0.56	0.16	0.21 / 0.59
In-bed leg movement	0.022	0.019	0.23	0.93	0.29	0.95 / 0.028
Adjusting bedding	0.037	0.038	0.25	0.93	0.29	0.23 / 0.52
Communication	0.240	0.257	0.27	0.93	0.29	0.080 / 0.20
Looking at others	0.330	0.320	0.74	0.93	0.74	0.54 / 1.0
Purposeful activity (any of four)	0.053	0.146	1.3×10^{-8}	–	–	5×10^{-7}

Table 15: **Cohort characteristics by delirium status.** Delirium is defined as any positive CAM-ICU assessment during the ICU stay. RASS percentages are computed over all recorded assessments in each group rather than over patients. Observed minutes run from arrival in the monitored room and can exceed ICU LOS, which counts only time at the critical-care level of care.

Characteristic	Delirium ($n = 53$)	Non-delirium ($n = 227$)
ICU LOS, days, median [IQR]	4.62 [3.01, 8.40]	3.09 [1.89, 4.84]
Observed minutes, median [IQR]	7,560 [4,680, 12,840]	4,920 [2,580, 7,440]
Invasive mechanical ventilation, n (%)	31 (58.5)	126 (55.5)
Per-patient median RASS, median [IQR]	0.00 [−1.00, 0.00]	0.00 [−0.50, 0.00]
RASS assessments ≤ -3, %	5.0 (of 6,979)	2.0 (of 18,402)

Table 16: **Delirium frequency ratios under weighting schemes with bootstrap intervals.** Minute-weighted ratios follow Eq. 1 and reproduce Figure 2(b); equal-weighted ratios average per-patient rates within each group. Intervals are 95% percentile bootstrap intervals over 10,000 patient-level resamples. Values below 1.00 ($\log_2$ below zero) indicate actions detected less frequently among patients with documented delirium. The final two actions are excluded from the 13-action evaluation set and from Figure 2(b) for lack of sufficient positive instances in the annotated benchmark.

	Minute-weighted		Equal patient weight	
Action	FR	$\log_2$ FR [95% CI]	FR	$\log_2$ FR [95% CI]
Using non-medical objects	0.41	-1.28 $[-1.92, -0.75]$	0.39	-1.35 $[-1.94, -0.87]$
Leg ADL movement	0.46	-1.13 $[-1.69, -0.67]$	0.47	-1.07 $[-1.59, -0.64]$
Object exchange	0.48	-1.05 $[-1.57, -0.61]$	0.47	-1.08 $[-1.55, -0.69]$
Gesticulating	0.60	-0.74 $[-1.17, -0.35]$	0.68	-0.56 $[-0.97, -0.21]$
Other arm movements	0.62	-0.70 $[-1.00, -0.42]$	0.69	-0.53 $[-0.79, -0.29]$
Grabbing medical equipment	0.67	-0.58 $[-0.87, -0.30]$	0.80	-0.33 $[-0.58, -0.09]$
Head turning	0.85	-0.24 $[-0.43, -0.07]$	0.90	-0.15 $[-0.30, -0.01]$
Adjusting bedding	0.85	-0.23 $[-0.58, 0.14]$	1.07	0.10 $[-0.31, 0.49]$
Position change	0.85	-0.24 $[-0.61, 0.14]$	1.06	0.09 $[-0.37, 0.52]$
Looking at others	0.96	-0.06 $[-0.25, 0.12]$	1.00	-0.00 $[-0.17, 0.15]$
In-bed leg movement	1.01	0.01 $[-0.38, 0.42]$	1.30	0.38 $[-0.06, 0.82]$
Communication	0.88	-0.18 $[-0.42, 0.04]$	0.94	-0.10 $[-0.30, 0.09]$
Clinician–patient contact	1.20	0.26 $[0.14, 0.39]$	1.18	0.24 $[0.14, 0.35]$
Not in evaluation set (detection performance not validated)				
Other torso movement	0.90	-0.16 $[-0.46, 0.15]$	1.08	0.10 $[-0.25, 0.45]$
Clinician restraining	1.54	0.62 $[0.15, 1.07]$	1.33	0.41 $[0.01, 0.79]$

A.15 Ethics, Data, and Code Availability

Human subjects review. This study was approved by the Institutional Review Board of the study institution. All video recordings and linked electronic health record data were stored and processed on secure servers, with access restricted to IRB-approved study personnel.

Data and code availability. The continuous video recordings and linked electronic health records used in this study contain protected health information and cannot be made publicly available due to patient privacy regulations and the terms of our IRB approval. Model prompts, the atomic action taxonomy, inference code, and evaluation code are available at `https://anonymous.4open.science/r/tracer-release-03B2`.