EGOEASY ANNOTATOR: EVIDENCE-GOVERNED TASK-NAVIGATION TIMELINES FROM IN-THE-WILD EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Human egocentric video could scale robot learning, but an untrimmed recording is not a ready-made training example. Recovering tasks, operations, and movement between work areas requires reading fleeting or obscured clues without inventing details or missing events. We introduce EGOEASY ANNOTATOR (EEA), which checks proposed descriptions against the video and separately searches for missed events. When interpretations would change the annotation, it asks what the video must show to distinguish them and opens suitable new views. A video model describes these views without seeing the descriptions being checked. A controller retains a specific description when the evidence distinguishes it from the alternatives, or shared details when only those are supported. Unclear details that still affect the annotation receive local human review; unsupported guesses are omitted. Thus, uncertain object or work-area names need not erase supported actions or navigation. Each newly found event undergoes the same checks before entering one source-timestamped task–navigation timeline. Averaging HD-EPIC and EgoEasy-Bench scores, EEA improves episode segmentation F1 by 12.34 percentage points over the best baseline, while recovering more events with fewer unsupported claims. In a 36-video study, it saves 32.14% of human review and correction time over the development-selected baseline while meeting the preset requirements for average final quality across both benchmarks.

1 INTRODUCTION

Human egocentric video offers a way to scale robot-learning data beyond teleoperation-based collection. Such recordings could support Vision-Language-Action (VLA) models that connect visual observations and language to robot actions, and World Action Models (WAMs) that learn how scenes evolve under action (Kareer et al., 2025; Li et al., 2026a; Jiang et al., 2026; Shi et al., 2026). Yet a long first-person recording is not a ready-made training example. To represent the full activity, annotations must group actions serving one objective into task episodes and preserve movement between work areas as navigation. The missing interface is therefore a language-aligned timeline that connects task episodes through navigation. We ask whether this timeline can be recovered directly from one untrimmed first-person stream, without external views, task scripts, supplied questions, or video-specific event lists. Just as importantly, doing so should reduce the substantial human effort otherwise required to segment and annotate long raw recordings.

Consider a wearer picking up a partly occluded object, walking through a doorway, moving it against a surface, and later returning. Holding an object does not establish tool use; a short move toward another work area does not establish navigation; and a plausible final view does not establish task completion. Even a short return to a previous work area can be missed entirely. An annotation can therefore fail in two opposite ways: it can say more than the video supports, or leave out events that actually occurred. Existing methods address parts of this problem, but often already specify the actions, segments, or questions the system should focus on. Fixed-vocabulary segmentation assigns known action labels (Abu Farha & Gall, 2019); dense annotation enriches existing demonstration segments (Kim et al., 2026); and active video agents, including hypothesis-verification methods, seek evidence for supplied questions (Wang et al., 2024a; 2026a). Our annotator instead must discover which events matter and how specifically each can be described. Refinement can improve a draft (Madaan et al., 2023; Gou et al., 2024), but does not ensure that missed events are found; finding

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.

more events does not by itself verify their meaning. Reliable whole-video annotation therefore needs both targeted verification and a separate search for omissions.

This creates two complementary requirements. First, any proposed interpretation that could change the annotation must be resolved to the most specific description supported by visual evidence, rejected, or left for human review. We call this *claim closure*. Second, a separate pass checks for missed events. Every important finding must be added, shown to be already covered, or marked for human review. We call this *coverage closure*. The first checks what the annotation says; the second checks what it may have missed.

EGOEASY ANNOTATOR (EEA) meets these requirements through controlled visual investigation. It first uses a video model to record visible contact, movement, and state changes with timestamps, forming an initial annotation. When different interpretations would change that annotation, it asks what the video must show to distinguish them. For use versus carrying, it checks contact and motion against a target; for completion, it follows the action from its starting state to its intended result. If the current views leave the question open, EEA requests new frames or clips rather than treating repeated answers as new evidence. To avoid suggesting an answer, it asks the same model to describe these views in a separate call without showing the initial annotation or proposed labels. A rule-based controller then uses these observations to decide what can enter the annotation. It keeps a specific description only if the video supports it, does not contradict it, and rules out the other proposed interpretations. If several interpretations still fit, it retains only what they share and the video supports. If an unclear detail still affects the annotation, only that detail or video interval needs human review; unsupported guesses are excluded. Thus, supported actions and navigation remain even when the exact object type or work area is unclear. This is *useful honesty*: keep what the video supports without guessing from context.

Even after these checks, the annotation can still miss events. EEA therefore separately reviews observations across the video, opening new clips when needed. This check sees which observations are already in the timeline, but not its descriptions or proposed labels. Newly found events receive the same checks before being added. To decide where one task ends and another begins, it checks for a change of goal, not merely a tool change or pause. For robot training, navigation between work areas also matters. Navigation requires leaving one work area, moving between areas, and stable arrival in another, not merely stepping toward it. Outward and return trips are two navigation episodes, even when the work areas remain unknown. EEA assembles one gap-free, non-overlapping timeline with the original video timestamps. Final descriptions use only details the controller approved, not guesses removed during checking.

We compare EEA with five baselines on 488 videos (51.817 hours) from HD-EPIC and EgoEasy-Bench. All use the same backbone and can inspect the full videos. We average the two benchmark scores and report gains over each metric's best baseline. Segmentation F1 improves by 12.34 percentage points (pp). Important-event recall rises by 13.52 pp, counting only events the evaluator confirms were correctly identified. The rate of unsupported core claims falls by 7.67 pp, including claims the evaluator cannot verify. EEA segments and annotates 13.58% of human-marked episodes without needing changes to boundaries or main content, versus the best baseline's 3.37%. Ablations show that claim closure reduces unsupported descriptions and coverage closure recovers missed events. In a 36-video study, EEA saves 32.14% of human review and correction time over the baseline selected on development videos. Average final quality across both benchmarks meets the preset requirements.

Our contributions are:

- We formulate annotation of untrimmed egocentric video around two requirements: say only what the video supports (*claim closure*) and recover important events that were missed (*coverage closure*).
- We introduce EGOEASY ANNOTATOR, which gathers new visual evidence for ambiguities that would change the annotation, checks that evidence without revealing the proposed answer, and separately searches for missed events.
- We produce a complete task–navigation timeline directly from untrimmed egocentric video, rather than answering supplied questions or annotating presegmented clips, and substantially reduce the human correction needed afterward without any drop in final annotation quality.

2 Related Work

Egocentric video and language for robot learning. Human first-person video is becoming a practical source of robot-learning data. EgoMimic combines human and robot demonstrations, EgoWAM studies world-action learning from in-the-wild human video, and recent whole-body systems use egocentric demonstrations for loco-manipulation Kareer et al. (2025); Li et al. (2026a); Jiang et al. (2026); Shi et al. (2026). DeMiAn adds multi-aspect language to existing robot and human-egocentric demonstration segments and improves both VLA and WAM learning Kim et al. (2026). EgoSmith, introduced with EgoSteer, filters locomotion and reconstructs hand and camera motion to curate manipulation-focused training data Zhong et al. (2026). Open-AoE releases natural recordings with text, hand poses, camera trajectories, and temporally localized atomic actions Li et al. (2026c). These works establish the value of scalable curation and language. EEA addresses a complementary whole-video decision: discover task structure and wearer transitions from one untrimmed ego stream, then release each semantic field only at the precision supported by that stream. In particular, navigation is supervision to preserve rather than background motion to discard.

Egocentric datasets and temporal annotation. EPIC-KITCHENS-100 and Ego4D provide large-scale egocentric actions and activities; Ego4D Goal-Step adds hierarchical goals and steps; HD-EPIC supplies dense multimodal labels; and Ego-Exo4D adds synchronized external views Damen et al. (2022); Grauman et al. (2022); Song et al. (2023); Perrett et al. (2025); Grauman et al. (2024). Dense captioning and temporal grounding methods scale localized descriptions or query-conditioned retrieval Yang et al. (2024); Zeng et al. (2025); Ren et al. (2024); Hannan et al. (2025), while temporal action segmentation assigns frames under a predefined action set Abu Farha & Gall (2019). These resources and methods provide valuable narrations, sensors, temporal queries, or fixed taxonomies. EEA studies a different inference contract: only the first-person stream is available, and the video-specific events, boundaries, roles, states, and navigation occurrences are not supplied.

Active evidence seeking in long video. Long-video agents improve reasoning by building memories, retrieving clips, changing temporal scale, or planning new observations Wang et al. (2024a); Fan et al. (2024); Ma et al. (2025); Li et al. (2026b); Wang et al. (2026b); Yang et al. (2026b). VideoHV-Agent turns candidate answers into hypotheses and discriminating visual clues, while REVEAL checks whether retrieved evidence satisfies a question-specific rubric and retrieves missing clues Wang et al. (2026a); Yan et al. (2026). A4VL and SVAgent coordinate multiple perception and reasoning roles Xu et al. (2026); Yang et al. (2026a). These methods primarily answer supplied questions, so the query or candidate answers define what to inspect and when the search is complete. Whole-video annotation must first create the output-changing questions and must also search outside its current hypotheses for events that were never proposed. EEA further separates semantic proposal from visual observation: the observer receives the relation to inspect but not the competing labels, and repeated reasoning over unchanged pixels adds no support.

Reliable annotation and selective review. Model-assisted annotation and tool-based critique show how language models can accelerate labeling and improve drafts Madaan et al. (2023); Gou et al. (2024); Wang et al. (2024b); Lin et al. (2025), while VidHalluc documents action, temporal-order, and scene-transition hallucinations in video models Li et al. (2025). SPARC produces structured spatial annotations from robot demonstrations and attaches reliability scores that improve the precision–coverage trade-off Blank et al. (2026). Selective prediction similarly formalizes abstention when risk is high Geifman & El-Yaniv (2019). EEA addresses a different annotation unit and failure set. From ego video alone, it checks roles, states, task boundaries, and navigation against the video, while preserving a supported broad value when an exact one is unsafe. It also prevents an exact guess from surviving behind a review flag and adds a separate coverage obligation for missed events.

3 EgoEasy Annotator

EEA checks both what an annotation says and what it leaves out. It first builds an initial annotation, investigates descriptions that the video may not support, and separately looks for missed events (Fig. 1). Both paths use the same rules to decide what to keep before producing one task–navigation timeline.

3.1 From video to an initial annotation

The input is one untrimmed egocentric video with N frames, $V = \{(x_i, \tau_i)\}_{i=1}^{N}$, where x_i is a frame and τ_i its original timestamp. No external view, task script, prior annotation, question, or event list is

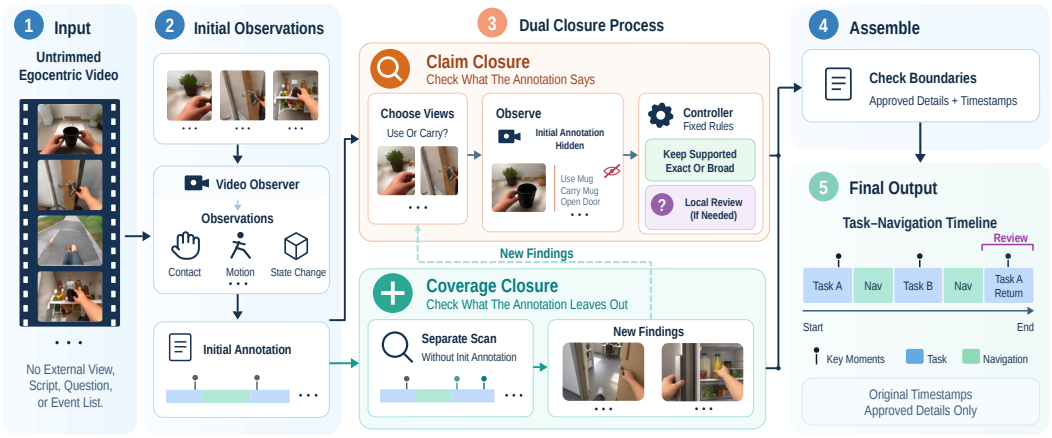

Figure 1: **Overview of EGOEASY ANNOTATOR.** Initial video observations form a tentative annotation. *Claim closure* checks uncertain descriptions using new views, with candidate labels and the initial annotation hidden from the observer. *Coverage closure* separately searches for missed events, which undergo the same checks. A rule-based controller retains supported exact or broader descriptions and flags important uncertainty for local review. Boundary checks and assembly produce a task–navigation timeline with key moments and original timestamps.

supplied. Let t_0 and t_1 be the video's start and end times. The output contains K episodes. Let b_k be the start of episode k and b_{K+1} the video end. Episode k occupies the interval I_k, with

$$I_k = [b_k, b_{k+1}), \quad k = 1, \ldots, K, \qquad t_0 = b_1 < \cdots < b_{K+1} = t_1, \tag{1}$$

so every time in $[t_0, t_1)$ belongs to exactly one episode. A TASK groups actions serving one goal, including preparation, checks, repetitions, and cleanup. NAVIGATION records movement between work areas. CANNOTIDENTIFY marks a span whose activity remains unclear, not an otherwise clear activity with an unknown object name. Observed rest or waiting is kept as a TASK record with an explicit background description. Important operations within an episode are recorded as timestamped *key moments*, rather than automatically becoming separate tasks.

A video model first examines views distributed across the whole video and records visible contact, movement, and state changes with their timestamps. During this first pass, it does not infer tool use, completion, work-area names, or object identity across an unseen interval. In another call, the model proposes episodes and descriptions from these observations, then considers other interpretations. We call each proposed detail tied to a time interval a *claim*, such as "using the object as a tool." For each claim, EEA records its description, time span, and the observations used to check it. EEA checks differences that would change an action, boundary, or need for review; paraphrases alone do not require investigation.

3.2 TURN UNCERTAIN DESCRIPTIONS INTO VISUAL CHECKS

For each such difference, EEA asks what the video must show to distinguish the interpretations. Use versus carrying calls for inspecting how an object contacts and moves against its target. For a completion claim, it inspects the state before the action, the action itself, and whether the resulting state meets the proposed task goal. Seeing a final state alone does not establish how it was reached. The model requests a frame, crop, or clip suited to this question. A still may show contact, whereas movement and state change require views over time. Tools retrieve these views and their original timestamps, not answers to the question.

To avoid suggesting an answer, EEA asks the same model to inspect these views for the requested relation in a separate call without candidate labels or the initial annotation. When the relation is visible, the observation is PRESENT; only a view sufficient to rule it out permits ABSENT. If the view includes the relevant action but remains unclear, the result is AMBIGUOUS; if the needed time or region was not shown, it is NOTVISIBLE. Neither means the event did not occur. A rule-based controller checks these observations against requirements fixed for each kind of claim; the model can request more checks, but cannot weaken them. The controller records the interpretations and views already inspected. After each observation, the model proposes a revised claim or requests another

view to settle the remaining question. A new request must expose relevant new pixels; repeated views or agreeing answers add no visual evidence. Appendix A.1 gives the claim records and evidence requirements; Appendix A.2 details view selection and the isolated calls.

3.3 Decide what the annotation can say

These checks apply to every claim, even when no competing interpretation was proposed. The controller keeps an *exact* description only if the video supports the required relations, checks find no evidence against it, and the other proposed interpretations are ruled out. Not disproving a guess is insufficient. If an exact description remains unsupported, EEA looks for supported information the possible interpretations share. Predefined mappings link specific labels to broader descriptions. Let H denote the remaining interpretations. For each $h \in H$, let $G(h)$ contain the descriptions permitted by these links. A shared description b is allowed only if

$$b \in \bigcap_{h \in H} G(h) \qquad \text{and} \qquad \text{the video supports } b. \tag{2}$$

The controller selects the most specific shared description using these links, then checks its visual support before keeping it as *broad*. Without a supported choice, the detail stays unknown; a claim contradicted by the video is removed.

An uncertain object type need not prevent recording visible tool use. But uncertainty about use versus carrying changes the action, so it requires human review of that detail or interval. Supported shared information remains, but the uncertain exact detail is withheld. Review never gives permission to keep an unsupported guess. If the activity itself cannot be established, the interval becomes CannotIdentify. Investigation stops when these decisions are settled or the preset budget is exhausted. Reaching the limit preserves supported information and sends remaining important questions to local review, rather than choosing the likeliest guess. This completes *claim closure* while preserving *useful honesty*. Appendix A.3 specifies decisions and stopping limits.

3.4 Find missed events and check episode boundaries

Even after these checks, the annotation may still miss an operation inside a long task episode. EEA therefore uses a separate model call to inspect observations from across the video and request new views where an event may have been missed. This call sees which observations and intervals are already covered, but not the annotation's descriptions or proposed labels. New findings return to the same visual checks, rather than entering the annotation as facts. Findings that could change an episode, key moment, boundary, or review decision cannot be ignored. The system links each one to a checked episode or key moment, explains with evidence that it is already covered, or marks its time and question for human review. This is *coverage closure*: every such finding has an outcome, though the scan may still miss events.

Finding the events still leaves the question of how to group them. To distinguish a task change from an internal step, EEA inspects activity before and after the proposed boundary. A new task requires evidence that the earlier goal ended or was interrupted and a different goal began and persisted. If one activity resumes after another, EEA checks both change points and whether the original work was left unfinished. Keeping the intervening activity within that task requires evidence linking them, such as a visible handoff or state change. A tool change or pause alone justifies neither a split nor a merge. If neither grouping is supported, the boundary remains for review.

Navigation requires leaving one work area, moving between areas, and stable arrival in another. Stepping toward another area or moving an object does not suffice. A claim about crossing a specific doorway or separator additionally requires seeing that crossing. If the video shows movement between work areas but not that passage, EEA records the movement without claiming that crossing. Work-area names are checked separately from the movement itself. For work areas A and B, a supported $A \to B \to A$ route produces two navigation episodes, even when the areas cannot be named. Appendix A.4 expands the coverage records, interrupted-task checks, and navigation cases.

3.5 Assemble and validate the final timeline

Once every claim and coverage finding has a recorded decision, including requests for human review, EEA rebuilds the timeline from the approved details, timestamps, and evidence links. A recovered event can split an earlier episode; descriptions retained in the resulting pieces must still have evidence within those pieces. Final text is generated from approved details, not by editing the initial annotation,

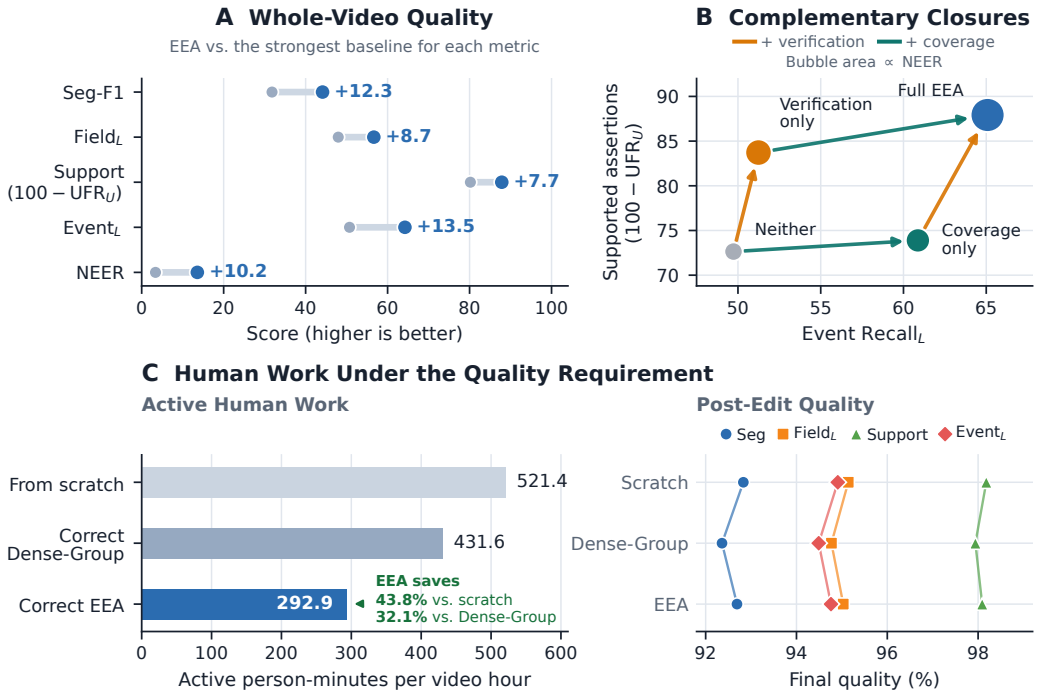

Figure 2: **Better timelines, complementary closures, and less human work.** Results equally weight HD-EPIC and EEB; L/U denote semantic lower/upper bounds, and Support is $100 - \text{UFR}_U$, a conservative supported-claim percentage. **(A)** EEA (blue) versus the strongest baseline per metric (gray), with gains in percentage points. **(B)** Arrows compare independently run conditions with verification (orange) or coverage (teal) added; bubble area indicates NEER. **(C)** Human work (left) and post-edit quality (right) for scratch annotation and correction of Dense-Group or EEA outputs. EEA reduces work within all eight preset limits on final quality averaged across benchmarks.

so discarded guesses are not copied back. The system checks Eq. 1 and that every key moment lies within its episode. If the final text conflicts with the approved claims, the affected claim must be checked again before release. The text check itself cannot add visual evidence or approve a more specific description. The system discards text that fails these checks and keeps the last validated timeline; without one, it records a failed run rather than an uncertain activity. Appendix A.5 gives the rebuilding and final-check procedure.

4 Experiments

Claim closure checks what the annotation says; coverage closure looks for missed events. We test whether annotations improve, which mechanisms contribute, and how much human work remains (Fig. 2).

4.1 Experimental setup

Videos and references. EgoEasy-Bench (EEB) tests task–navigation timelines, required facts, and operations in cleaning, food-preparation, and organization videos. Most valid files last five minutes, but 53.83% of navigation episodes last at most five seconds: short trips must be retained within longer activities. We test 428 held-out EEB files (34.806 hours) and 60 HD-EPIC recordings (Perrett et al., 2025) (17.010 hours), totaling 488 files and 51.817 hours; HD-EPIC extends to 44 minutes. Files are neither trimmed nor concatenated. Three annotator groups cross-check reference episodes, necessary facts, and operations without seeing predictions. EEA was not fine-tuned or behaviorally adjusted using these videos. Appendices B.1 and B.2 describe the collection, references, and development/test separation.

Comparable systems. All generation calls use BF16 Qwen3.8-27B, ego RGB, original timestamps, and common output definitions; no method receives reference answers, supplied questions, event lists, or other modalities. All may retain unknowns and request review. Direct combines annotations

from uniformly sampled windows. Dense-Group combines descriptions from overlapping windows into episodes; Dense-Group+SR revises that annotation using only previously seen frames. DeMiAn-Ann (Kim et al., 2026) generates segments before multi-aspect annotation; VideoHV-Ann (Wang et al., 2026a) generates questions before hypothesis verification. These are annotation adaptations, not the methods' original-task results. We select baseline settings and the human-study comparator on development videos. Appendix B.3 gives settings and adaptations; Appendix A.8 provides their prompts.

Accuracy, completeness, and support. We score timing, required content, and visual support separately. Temporal intersection-over-union (tIoU) divides shared duration by the duration covered by either interval. Seg-F1 matches Task/Navigation episodes one-to-one by type and tIoU ≥ 0.5, excluding background. Field-F1 measures necessary facts: matches require supported meaning, the correct functional role, and tIoU ≥ 0.5. We average F1 across action, object/tool-role, state, and work-area categories. EEB lacks exhaustive state references, so its state claims undergo support checks but not this category average. Event Recall measures recovery of distinct reference operations by timestamped key moments. A match requires the same operation and occurrence. The time can be within one second of a human timestamp. Otherwise, it must lie inside the reference episode (one-second edge tolerance), with video frames within one second of the prediction confirming the operation.

A fixed Qwen3.8-27B evaluator extracts required facts/operations from human references, then checks anonymous annotations against references and frames. Programs count and calculate scores; the evaluator never assigns an overall score. Let A count distinct nonempty core claims, B those contradicted or insufficiently supported, and U those still undecided. Unsupported Field Rate (UFR) lies in $[B/A, (B+U)/A]$; decorative details are excluded. These bounds reflect undecided judgments, not sampling uncertainty. Field/Event lower bounds count only confirmed matches; UFR's upper bound includes undecided claims. Human and evaluator labels agree on 218/240 sampled judgments (90.83%); Appendix C.5 reports disagreements and scorer limitations.

A correct but incomplete episode is not ready to use. No-edit episode recall (NEER) is the fraction of reference Task/Navigation episodes whose matched prediction needs no core corrections. That prediction must recover every required fact/event, place both boundaries within one second, and have no unsupported or undecided core claim or substantive review request. Review span is the fraction of video time marked for core review, not human time. Failed files count as entirely needing review and as missing their reference episodes, facts, and events. Appendix C.2 specifies matching and empty denominators.

We pool counts within each benchmark, compute its scores, then average the two benchmarks. Confidence intervals use 10,000 paired bootstrap resamples of whole groups (7 HD-EPIC wearers, 11 EEB collection groups), not individual files. The primary endpoints are Seg-F1, Field lower bound, and UFR upper bound; their 15 baseline comparisons use group-level exchange tests with Holm correction (Appendix C.4).

4.2 Better timelines without discarding useful information

Recover more, with fewer unsupported claims. EEA improves episode recovery and the reliability of descriptions (Table 1). Against the best baseline for each primary metric, Seg improves by 12.34 percentage points (pp; 95% CI $[6.08, 18.57]$), Field by 8.67 pp $[1.47, 14.85]$, and UFR falls by 7.67 pp $[2.04, 13.27]$. All 15 Holm-adjusted tests have $p < 0.05$; each benchmark shows all three improvements (Appendix D.1).

Table 1: Results averaged across benchmarks. Quality/review values are percentages; L/U are bounds from undecided judgments. Tokens are millions per video hour, including visual input, output, repairs, and failed attempts.

Method	Seg↑	Field$_L$ ↑	UFR$_U$ ↓	Event$_L$ ↑	NEER↑	Review	Tokens
Direct	19.55	32.75	30.16	33.55	0.78	26.52	0.90
Dense-Group	31.80	46.14	21.68	50.29	2.86	14.12	1.87
Dense-Group+SR	29.84	47.58	19.80	50.71	3.37	16.12	5.78
DeMiAn-Ann	17.64	36.81	32.49	38.17	0.88	15.02	4.47
VideoHV-Ann	27.63	47.96	24.23	49.09	2.86	21.51	13.39
EEA	**44.13**	**56.64**	**12.14**	**64.23**	**13.58**	22.14	8.98

Lower UFR does not come from saying less: Event Recall rises, and NEER increases from 3.37% to 13.58%. Nor does EEA achieve this by making every description vague. For required facts the video can establish precisely, Observable Exact Recall measures recovery at that precision. It rises from the best baseline's 46.67% to 52.64%. EEA's UFR upper bound is below every baseline's lower bound, so fewer undecided judgments alone cannot explain its advantage. Appendix D.2 checks judgment and sampling uncertainty.

Preserve trips between work areas. Better navigation should mean fewer missed trips, not more motion labeled as navigation. Navigation F1 rises from Dense-Group's 20.24% to 37.73%; recall of complete round trips rises from 12.86% to 34.08%, requiring both legs. We also test premarked camera-turn, object-motion, and local-step intervals: navigation covering at least half an interval counts as a false detection. This rate falls from the best baseline's 22.50% to 10.30%. Thus, EEA retains more trips while rejecting more local motion; Appendix D.3 adds boundary and endpoint results.

These gains cost 4.79 times Dense-Group's tokens, but 0.67 times VideoHV-Ann's. Denser Dense-Group trails EEA while using fewer tokens; this is not equal compute (Appendix D.4). EEA completes 98.35% of runs after averaging benchmarks; failures remain scored. The improvements are not confined to the main backbone: replacing every generation call with Qwen3-VL-30B-A3B-Instruct on 60 files retains +11.54 pp Seg, +5.63 pp Field, and -8.53 pp UFR over VideoHV-Ann (Appendix G).

4.3 Which mechanisms account for the gains?

To separate checking claims from finding omissions, we switch verification with release checks and coverage independently on or off on 60 preselected files, 30 per benchmark (Table 2). All four combinations start from identical observations and annotations, then run independently without Full EEA's later evidence. Component controls keep Full's investigation locations and order but recompute decisions. Uniform views uses equally many new views from the same allowed ranges. Labels visible shows proposed answers with unchanged views. No broad outputs sends information supported only at a broader level to review, without weakening exact-label checks.

Table 2: Mechanism results on 60 files. Quality/review values are percentages. Cost is processed tokens relative to Full. The lower block changes one component along Full's fixed investigation schedule.

Condition	$\text{UFR}_U \downarrow$	$\text{Event}_L \uparrow$	NEER↑	Review	Cost
Neither path	27.36	49.73	4.18	25.37	0.472
Verification only	16.28	51.24	8.61	29.48	0.714
Coverage only	26.11	60.87	7.22	28.73	0.703
Full	12.08	65.08	14.21	21.86	1.000
Uniform views	16.19	60.13	10.46	26.48	1.019
Labels visible	17.31	62.51	11.62	23.71	0.993
No broad outputs	11.14	59.14	10.31	37.64	0.971

Checking claims does not replace finding omissions. Verification alone lowers UFR by 11.08 pp, but its 1.51 pp Event Recall gain is uncertain (CI $[-3.57, 6.59]$). Adding coverage raises Event Recall by 13.84 pp $[8.58, 19.10]$ and NEER by 5.60 pp. Conversely, adding verification to coverage lowers UFR by 14.03 pp. Thus, checking claims improves support, while searching separately recovers omissions (Appendix E.1).

What to inspect, and what to release. Choosing the right views matters: Full lowers UFR by 4.11 pp versus equally many uniform views. Replacement is possible for 783/840 investigations; the rest stay unchanged. With evidence held fixed, hiding proposed labels raises supported, correct choices from 60.40% to 73.20% among 250 initially wrong candidates. Initially correct candidates show no clear difference, so the benefit is correcting wrong suggestions. Retaining supported broad descriptions matters too: withholding them loses 5.94 pp Event Recall and adds 15.78 pp review, without clear UFR improvement. This control increases review by design, so usefulness needs an independent check. Humans confirm 55/60 withheld descriptions as supported and useful (three errors, two undecided): broad descriptions preserve useful information. Appendix E.2 details the controls and intervals.

4.4 FROM BETTER ANNOTATIONS TO LESS HUMAN WORK

NEER cannot tell us how long people spend finding and fixing remaining problems. We therefore compare scratch annotation with correction of Dense-Group or EEA on 36 files, 18 per benchmark, recording time and final quality (Table 3). Thirty-six operators complete 108 assignments with identical instructions and interfaces. None sees references or handles one video in different workflows. Time includes whole-video checking, editing, two independent production reviews, disagreement resolution, and rework (Appendix F.1).

Table 3: Human time and final quality, averaged across benchmarks. H averages the two benchmarks' total active person-minutes per video hour. All assignments finish; acceptance does not imply perfect scores.

Workflow	$H \downarrow$	Seg↑	Field$_L$ ↑	UFR$_U$ ↓	Event$_L$ ↑
From scratch	521.4	92.83	95.14	1.82	94.91
Correct Dense-Group	431.6	92.36	94.77	2.06	94.49
Correct EEA	**292.9**	92.69	95.03	1.91	94.76

Faster annotation must still meet the quality requirements. Before testing, we require 95% difference intervals to exclude losses of 2 pp or more in final Seg/Field/Event and increases of 1 pp or more in UFR. All four requirements pass against both controls, even with undecided judgments least favorable to EEA. Within these limits on average quality, EEA saves 43.82% $[36.9, 50.7]$ of human work versus scratch and 32.14% $[23.8, 40.4]$ versus Dense-Group. Not all individual-benchmark changes meet these limits; Appendix F.2 reports them and the full tests.

Independent checking finds 81 substantive problems in EEA's initial annotations versus 218 in Dense-Group's; only 20 versus 171 lie outside marked review areas. Thus, EEA leaves fewer problems to fix and fewer to discover unaided, consistent with lower human time despite its larger review span (Appendix F.3).

5 CONCLUSION

Turning untrimmed egocentric video into a useful task–navigation timeline requires both checking proposed descriptions and finding missed events. We introduced EGOEASY ANNOTATOR, which combines claim closure and coverage closure to build that timeline from the video alone. It gathers new evidence for uncertain interpretations, checks it without revealing the proposed answers, and applies the same rules to independently discovered events. Its controller keeps only the detail supported by the video and marks unresolved questions for review when they affect the annotation. This preserves supported actions and navigation even when finer details remain unclear, rather than filling gaps with guesses.

On HD-EPIC and EgoEasy-Bench, EEA recovers more important events and observable detail while reducing unsupported claims. The gains therefore reflect both better-supported content and more complete recovery, not simply a decision to say less. Ablations show why both checks matter: checking claims reduces unsupported descriptions, while searching separately recovers events missing from the initial annotation. In the human study, reviewing and correcting these annotations also takes less time. After correction, average quality across both benchmarks meets the preset requirements. Together, these findings point to a practical way to scale robot-learning data: not by removing human judgment, but by recovering more supported information and reducing the work needed to turn raw video into useful annotations.

AI USE STATEMENT

Generative AI tools were used solely for language-level editing, including checks and improvements of grammar, wording, clarity, and readability, as well as for presentation-level adjustments to the layout and scaling of figures and tables. These tools were not used to develop or modify the research methodology, formulate hypotheses or mathematical claims, design experiments, implement methods, process or analyze data, or interpret experimental results. The underlying scientific content, data, results, and conclusions were not altered by AI-assisted editing.

All AI-assisted modifications were reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text and visual materials.

REFERENCES

Yazan Abu Farha and Jürgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019.

Nils Blank, Paul Mattes, Maximilian Xiling Li, Jakub Suliga, Thomas Roth, Moritz Reuss, Pankhuri Vanjani, and Rudolf Lioutikov. SPARC: Reliable spatial annotations from robot demonstrations at scale. *arXiv preprint arXiv:2606.13497*, 2026. doi: 10.48550/arXiv.2606.13497. URL https://arxiv.org/abs/2606.13497.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 130:33–55, 2022.

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 *Computer Vision – ECCV 2024*, volume 15080 of *Lecture Notes in Computer Science*, pp. 75–92. Springer, 2024. doi: 10.1007/978-3-031-72670-5_5. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/3241_ECCV_2024_paper.php.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, pp. 57734–57811, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fef126561bbf9d4467dbb8d27334b8fe-Abstract-Conference.html.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19012–19022, 2025.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. In *International Conference on Learning Representations*, pp. 157438–157461, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ff7373914a96956f2a7cacbdf3b0b8d8-Abstract-Conference.html.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127989. URL https://doi.org/10.1109/ICRA55743.2025.11127989.

Bosung Kim, Ruiyi Wang, David Acuna, Jaehun Jung, Alexander Trevithick, Brandon Cui, Yejin Choi, and Prithviraj Ammanabrolu. How to instruct your robot: Dense language annotations power robot policy learning. *arXiv preprint arXiv:2605.17077*, 2026. doi: 10.48550/arXiv.2605.17077. URL https://arxiv.org/abs/2605.17077.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. *arXiv preprint arXiv:2607.08436*, 2026a. URL https://arxiv.org/abs/2607.08436.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_VidHalluc_Evaluating_Temporal_Hallucinations_in_Multimodal_Large_Language_Models_for_CVPR_2025_paper.html.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. LensWalk: Agentic video understanding by planning how you see in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19518–19528, 2026b.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, Zitong Shan, Zhenchao Jin, Jiadong Hong, Taowen Wang, Yushi Feng, You Liu, Yibo Wang, Yifan Yang, Zhaowen Zhou, Man Luo, Hao Cheng, Bo Zhang, Jianshu

Li, Jiansheng Cai, Guocai Yao, Jize Zhang, Chenhao Lin, Renjing Xu, Lequan Yu, Chao Shen, Chunhua Shen, and Zhe Li. Open-AoE: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026c. doi: 10.48550/arXiv.2607.14183. URL https://arxiv.org/abs/2607.14183.

Lequan Lin, Dai Shi, Andi Han, Feng Chen, Qiuzheng Chen, Jiawen Li, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, and Junbin Gao. ACT as human: Multimodal large language model data annotation with critical thinking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24107–24138. Curran Associates, Inc., 2025. doi: 10.52202/085713-0814. URL https://papers.nips.cc/paper_files/paper/2025/hash/22bbfc36cb32d0baf7cd1c832da6f5e9-Abstract-Conference.html.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Kumar Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23901–23913, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024.

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Tianyu Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. In *Proceedings of Robotics: Science and Systems*, 2026. URL https://www.roboticsproceedings.org/rss22/p204.pdf.

Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4D goal-step: Toward hierarchical understanding of procedural activities. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, 2024a.

Yifan Wang, David Stevens, Pranay Shah, Wenwen Jiang, Miao Liu, Xu Chen, Robert Kuo, Na Li, Boying Gong, Daniel Lee, Jiabo Hu, Ning Zhang, and Bob Kamma. Model-in-the-loop (MILO): Accelerating multimodal AI data annotation with LLMs. *arXiv preprint arXiv:2409.10702*, 2024b. URL https://arxiv.org/abs/2409.10702.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. *arXiv preprint arXiv:2603.04977*, 2026a. URL https://arxiv.org/abs/2603.04977.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9088–9099, 2026b.

Yichang Xu, Gaowen Liu, Ramana Rao Kompella, Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, and Ling Liu. A multi-agent perception-action alliance for efficient long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19497–19507, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Xu_A_Multi-Agent_Perception-Action_Alliance_for_Efficient_Long_Video_Reasoning_CVPR_2026_paper.html.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. REVEAL: A rubric-guided agent for explicit evidence sufficiency verification in long-video question answering. *arXiv preprint arXiv:2608.08612*, 2026. doi: 10.48550/arXiv.2608.08612. URL https://arxiv.org/abs/2608.08612.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In *Advances in Neural Information Processing Systems*, volume 37, pp. 57240–57261. Curran Associates, Inc., 2024. doi: 10.52202/079017-1824. URL https://papers.nips.cc/paper_files/paper/2024/hash/6903a5aaece71b76623245fc6e32f01b-Abstract-Datasets_and_Benchmarks_Track.html.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. SVAgent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24062–24072, 2026a.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, and Lidong Bing. LongVT: Incentivizing "thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_LongVT_Incentivizing_Thinking_with_Long_Videos_via_Native_Tool_Calling_CVPR_2026_paper.html.

Yingsen Zeng, Zepeng Huang, Yujie Zhong, Chengjian Feng, Jie Hu, Lin Ma, and Yang Liu. DisTime: Distribution-based time representation for video large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21961–21971, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zeng_DisTime_Distribution-based_Time_Representation_for_Video_Large_Language_Models_ICCV_2025_paper.html.

Yifan Zhong, Zhang Chen, Tianrui Guan, Fanlian Zeng, Yuyao Ye, Tianjia He, Ka Nam Lui, Jiayi Li, Tingrui Zhang, Ruilin Yan, Xinhao Ji, Guangyu Zhao, Wenjie Lou, Jiayuan Zhang, Yuanpei Chen, and Yaodong Yang. EgoSteer: A full-stack system towards steerable dexterous manipulation from egocentric videos. *arXiv preprint arXiv:2607.09701*, 2026. doi: 10.48550/arXiv.2607.09701. URL https://arxiv.org/abs/2607.09701.

A ADDITIONAL METHOD DETAILS

This appendix follows the annotation process in Section 3. We first specify the evidence needed to support a description, then explain how model calls obtain that evidence, handle uncertainty, and find missed events. We next describe how the accepted information becomes the final timeline. The last sections give the model settings and the complete prompts used for EEA and our baseline adaptations.

A.1 CLAIM RECORDS AND EVIDENCE REQUIREMENTS

Keep a proposed statement separate from its evidence. A claim represents one proposed detail over a time interval, such as an action, goal, object role, state, task boundary, navigation event, work area, or key moment. Its record identifies the source video and interval, the proposed description and its level of detail, the other interpretations being considered, and the observations used to check them. Times use the original video's presentation timestamps and time base, not frame indices from a resized preview. If the proposed meaning changes, the revised claim remains linked to the earlier one. This link explains the revision; it does not supply new visual evidence.

An inspected clip and an event observed within it need not have the same interval. Seeing an action somewhere in a clip does not show that it lasts throughout the clip. Each observation therefore records both the view inspected and the time over which it supports a claim. When an episode is split, the descriptions retained in each piece must still have evidence within that piece. The video's original duration determines the final endpoint in Eq. 1, rather than the timestamp of the last sampled frame.

Apply the same evidence requirements to every claim. Before accepting a description, the controller checks that the source, timestamps, response structure, and evidence references are valid. An exact description also needs direct visual support, a check for evidence against it, and exclusion of the other interpretations under consideration. Model confidence, agreement across calls, and the absence of a challenge cannot replace these checks. Table 4 specifies the additional relations required for each kind of description. The model may request further inspection but cannot remove a required check.

Table 4: Visual requirements for exact descriptions. Every row also requires direct support, a check for evidence against the claim, and exclusion of the competing interpretations.

Proposed detail	Additional evidence
Action descriptions	At least one group of relations appropriate to the action: (i) contact and relative motion; (ii) contact and visible use; (iii) identity continuity and relative motion; (iv) the before state, connecting action, after state, and state change; (v) the earlier goal ending or being interrupted, a different goal beginning, and its continuation; or (vi) departure, boundary, passage, and arrival. This applies to action labels, episode descriptions, instructions, key moments, and visible-evidence statements.
Task goal	Activity continuing toward that goal, together with one of groups (i)–(iv) or (vi) above. An ongoing goal need not be claimed complete.
Task boundary	The earlier goal ends or is interrupted; another goal begins and persists.
Tool role	Visible object structure, physical contact with the target, and visible use.
Target role	Visible object structure and at least one of physical contact, relative motion, visible use, or state change.
Work-area name	Visible scene structure and sufficient visibility to establish the area.
Before or after state	The corresponding before or after state is visible.
Task completion	The before state, connecting action, after state, and state change are all observed.
Fully specified navigation	Departure, passage across a visible boundary, arrival, stable origin and destination, and support for any endpoint labels. Unknown names do not erase independently supported movement.

The evidence must support the action actually being described. For example, contact can establish touching without establishing cleaning or task completion. Different sentences expressing the same action remain subject to the same evidence requirements. A claim with no specified annotation detail cannot receive an exact value.

A.2 Choosing views and separating model calls

Let the question determine the view. A viewing request identifies the claim, the relation that would distinguish its interpretations, the permitted time or region, and the views already inspected. The planner chooses one next visual action. A frame or crop can show static structure, whereas contact onset, release, and relative motion require temporal context. Completion requires the action together with its before and after states. Checking whether an object remains the same may require views across a gap or a put-down and pickup sequence. Task and navigation questions may require wider views around the proposed change, including earlier or later activity beyond the initial local view.

The tools return pixels, original times, and references identifying the retrieved views. They do not answer whether an object is being used, a task is complete, or the wearer has changed work areas. The controller records what has already been inspected. Repeating a view or receiving another agreeing answer cannot increase visual support. If a request would add no relevant new pixels, the planner either changes the request within its remaining allowance or stops under Appendix A.3.

Separate information, not model weights. The same backbone performs the jobs in Table 5 through calls with different inputs. Initial construction and later revision use the same prompt, with the call mode selecting the applicable instructions. The call that observes new views receives a neutral question about the visible relation, without the candidate labels or initial annotation. The revision and decision calls can then compare those observations with the candidates. Revision proposes an updated claim or another needed view; a subsequent call recommends a decision. Only the controller can accept that recommendation after checking the evidence requirements. The complete original prompts appear in Appendix A.7.

Table 5: What each EEA call receives and returns. All generation calls exclude human reference annotations. The controller checks response structure and evidence links before using the returned content.

Call	Receives	Returns and authority limits
Initial observation	Source views, original times, and viewing scope.	Visible facts, their times and evidence references, and limits of visibility. No prior annotation or candidate labels.
Initial construction	Initial observations and the annotation format.	Proposed episodes and claims, assumptions, alternative interpretations, and evidence links. No reference answers.
Challenge	One selected claim, its time, and supporting observations.	An alternative that would change the annotation, the distinguishing relation, and what could refute the selected claim. No unrelated detail.
Planning	The challenge, allowed actions, inspected views, and remaining work allowance.	One view request with time or region, required new pixels, and the expected decision effect. No semantic answer.
Separate observation	New views and times, plus a neutral relation question.	Visible facts and object tracks; relation results, checks against the interpretation, and limitations. No candidate labels or initial narrative.
Revision	The selected claim, new observation, earlier claim records, and remaining allowance.	A revised claim and its connection to the earlier claim, a next relation to inspect, or a stopping recommendation. No acceptance authority.
Decision recommendation	Candidates, observations, evidence requirements, and allowed outcomes.	Retained and removed claims, evidence checks, remaining uncertainty, and local review needs. The controller validates the decision.
Coverage	Source observations, their times, and links to the existing timeline.	Already covered observations and unresolved findings with their times, visible basis, and suggested inspection. No annotation narrative or labels.

Call	Receives	Returns and authority limits
Final readout	Approved episodes, original times, evidence links, and review requests.	The final annotation, preserving unknown values and review. No new facts, labels, boundaries, or key moments.
Final text check	Candidate annotation, accepted and removed claims, evidence, times, and review requests.	A pass or issues with affected claims and requests to recheck or block. No images, new-event discovery, or permission to release.

Record the response needed by the next step. Each call returns a structured JSON response containing the information in Table 5. An observation must identify the views supporting it; a revised claim must identify what changed; a decision must identify the information retained and any question left for review. The controller checks these references and the response format before using the content. Video-specific views, observations, candidates, and budgets are supplied with each request rather than embedded as fixed answers in the prompts. All visual calls in the reported experiments use the full designated monocular view, with the settings in Appendix A.6.

Do not confuse an unseen event with an absent event. The observer reports PRESENT when the requested relation is visible. It reports ABSENT only when the inspected time and region are sufficient to rule the relation out. AMBIGUOUS means the relevant content is shown but cannot be resolved; NOTVISIBLE means the required time, phase, or region was not shown. Both leave the question open. The report can rule out an interpretation only from evidence actually shown to that call.

Choose which question to investigate next. Pending work is ordered as follows: source or response-format failures, omissions, navigation, task boundaries, object identity across time, factual claims, object roles, description specificity, completion or intent, and equivalent wording. An explicit priority assigned by the controller takes precedence. Equal priorities are resolved by creation order and then a stable work identifier. Equivalent wording is merged before visual investigation, so paraphrases do not create additional conflicts.

A use-versus-carry investigation. One recorded investigation covers 10–12 seconds of a source video. The initial claim describes using an object as a tool; the alternative is carrying it. The planner requests a new continuous clip from contact onset through release, including the target region. Without seeing either label, the observation call reports contact, relative motion, and visible use as PRESENT, and counterevidence as ABSENT. The controller accepts the specific tool-use description because those observations provide positive support, complete the counter-check, and exclude the carrying alternative. This sequence illustrates how the question leads to a view, the view to an observation, and the observation to a decision.

A.3 SHARED DESCRIPTIONS, REVIEW, AND STOPPING

Find a common description without inventing one. When an exact interpretation remains unsupported, the controller looks for a description permitted by all remaining interpretations. Table 6 gives all fixed links used for this lookup. A model can propose a broader description but cannot add a link or make an otherwise ineligible value acceptable. Table 7 gives the visual checks for every explicitly permitted value. Being listed in these tables does not itself establish support in a particular video.

Table 6: All fixed links from specific descriptions to broader descriptions. Each entry in the right column has a direct link to the description in the middle column. An intermediate-only value can be traversed but cannot itself be selected for output.

Detail	Broader description	Specific descriptions linked to it
Action	object manipulation	add material to a vessel; coil or bundle a flexible object; cut object; cut or peel object; fit a cover or liner around an object; fold or smooth textile; open package; place an object in a container; pour material into a receptacle; rearrange items within storage; remove a part from an object; retrieve items from storage; rinse or agitate contents in water; stir contents; transfer an object between containers; wash object
Action	object interaction	carry object; close a container or door; handle object; hold object; open a container or door; pick up object; put down object; reposition object; turn on appliance; use a handheld device
Action	clean the floor with a floor-cleaning tool	clean the floor with a mop
Action	object–surface interaction	clean surface; contact surface; rinse surface with water; spray surface; wipe surface
Action	hold object	lift object
Action	lift object (intermediate only)	lift objects
Boundary	task boundary	a particular task split
Completion	task state change	completed
Navigation	move between work areas	crossing between distinct spaces
Navigation	crossing between distinct spaces	crossing between named spaces; room transition
Tool	long-handled floor-cleaning tool	mop

The lookup stays within the same kind of claim, such as action, tool, navigation, or completion. Starting from each remaining interpretation $h \in H$ in Eq. 2, it follows the links toward broader descriptions and finds those common to all interpretations. Let $d(h, b)$ be the number of links in the shortest allowed path from h to a common description b. It first minimizes $\max_{h \in H} d(h, b)$, then $\sum_{h \in H} d(h, b)$. If several descriptions remain tied, it returns no unique choice rather than deciding by alphabetical order. The selected description must then pass its own visual checks. A failed lookup or a lack of visual support leaves the detail unknown or for local review; neither permits a guessed exact label.

For example, wiping a surface and contacting a surface can share the description "object–surface interaction." That description still requires visible contact and relative motion. Likewise, movement between work areas can be retained without claiming a specific room crossing, but it requires its own departure and arrival evidence. Values with no additional relation in Table 7 still require source-supported meaning; they are not exempt from the common checks.

Table 7: Visual checks for every explicitly permitted output value in the fixed mapping. Values with identical checks share a row. These requirements do not replace the common checks or make a value applicable to every video.

Permitted description	Required visual evidence
add material to a vessel	the before state; the connecting action; the after state; a visible state change; a directly observed relation; the receiving region
carry object	holding; physical contact; relative motion; transport of the held object
clean the floor with a floor-cleaning tool; object–surface interaction	physical contact; relative motion
close a container or door	the before state; the connecting action; the after state; a visible state change; the change to a closed configuration
coil or bundle a flexible object; fit a cover or liner around an object; fold or smooth textile	the before state; the connecting action; the after state; a visible state change; a directly observed relation; physical contact; relative motion
cut or peel object	a visible state change; a directly observed relation; physical contact; relative motion; visible use
hold object	holding
object interaction; object manipulation	a directly visible relation
object nearby	visible proximity
object present	visible presence in the scene
open a container or door; transfer an object between containers	the before state; the connecting action; the after state; a visible state change
open package	the before state; the connecting action; the after state; a visible state change; the change to an open configuration; a directly observed relation; physical contact; relative motion
pick up object	physical contact; relative motion; the pickup action; the pickup arrangement and movement
place an object in a container	a visible state change; a directly observed relation; physical contact; the placement arrangement and movement; the receiving region
pour material into a receptacle	the before state; the connecting action; the after state; a visible state change; a directly observed relation
put down object	a directly observed relation; physical contact; the put-down action; the placement arrangement and movement
remove a part from an object	the before state; the connecting action; the after state; a visible state change; a directly observed relation; physical contact; relative motion; the part separating from the object
retrieve items from storage	a directly observed relation; physical contact; relative motion; the pickup action; the pickup arrangement and movement
rinse or agitate contents in water; stir contents; wipe surface	a directly observed relation; physical contact; relative motion; visible use
rinse surface with water	a directly observed relation; relative motion; visible use
spray surface	a directly observed relation; visible use
turn on appliance	the before state; the connecting action; the after state; a visible state change; a directly observed relation; physical contact
use a handheld device	a directly observed relation; physical contact; visible use
task boundary; task state change	No additional relation beyond source-supported meaning and the common checks.

Permitted description	Required visual evidence
crossing between distinct spaces	departure; the boundary crossed; the passage itself; arrival
move between work areas	departure; a stable origin work area; arrival; a stable destination work area
long-handled floor-cleaning tool	visible structure; physical contact

Keep useful information while locating uncertainty. The controller can keep an exact description, keep a supported broader description, keep that broader description with local review, or leave an important detail unknown with review. If the activity itself is unclear, its span is CANNOTIDENTIFY; an uncertain object type alone does not require that episode type. Review records the question and the affected detail or time interval. The supported information remains, but a review mark never permits an unsupported exact guess.

The response represents one of these four outcomes or a run failure. The original prompt also uses "CLEAR/REJECT" to instruct removal of a contradicted detail. This removal is recorded separately rather than treated as another decision category. The response separately records removed claims and details that were cleared or broadened. Removing one detail neither deletes other supported information nor removes its video interval.

Stop without lowering the evidence standard. After a new observation, the revision call can propose a revised claim, request another distinguishing relation, or recommend stopping. Investigation ends when an allowed decision can be made, the remaining distinction no longer changes the usable annotation, or the allotted work is exhausted. At the limit, supported shared information stays and important unresolved questions go to local review. Invalid source references, unrecoverable response errors, and invalid timeline structures are recorded as run failures, not hidden as uncertain activities.

Scale the work allowance with video length. The source is divided into consecutive scheduling windows of at most 30 seconds and 900 native frames, including the tail. Let n be the number of windows and define

$$s = \max(1, \lceil n/24 \rceil) . \tag{3}$$

The integer s sets the limits in Table 8; it is not a unit of seconds. Initial observation covers all n windows at 2 frames per second, rather than stopping at the first 24. Possible changes receive priority in the viewing order, but the other windows are not skipped. Assigning every native frame to a scheduling window does not mean that the model inspects every frame.

Table 8: Maximum work allowed for a video. These are safety limits, not amounts that every run must consume.

Operation	Limit
Further evidence requests	$48s$
Claim investigations	$24s$
Evidence requests per investigation	2
Controller steps	$256s$
Coverage rounds	2
Evidence requests per coverage round	$4s$
Final-check recovery rounds	2
Post-assembly recovery attempts	2
Calls per role	$n + 256s + 8$
End-to-end wall time	$7200 \times s$ seconds
Single-call transport timeout	7200 seconds

Initial observation and later evidence requests are counted separately. A source-change scan uses 64×64 images with darkness threshold 32 and change threshold 24 to identify possible viewing needs, not to assign activity labels. The limits depend on source length, not reference annotations or quality scores.

A.4 COVERAGE, INTERRUPTED TASKS, AND NAVIGATION

Give each important finding a place in the annotation. The coverage call receives observations from across the video, their intervals, and links showing which observations are already represented by episodes or key moments. It cannot read the initial annotation's wording or candidate labels. It checks for missing internal operations, changes of goal, outward and return movement, and weakly observed intervals concealed by a neighboring description. A viewing-window edge is not automatically a task boundary: one operation can span windows, and one window can contain several operations.

A finding matters if ignoring it could change an episode, key moment, boundary, navigation event, or review decision. Let $\mathcal{O}_{\mathrm{mat}}$ contain the findings exposed by the scan that matter in this sense, and let $d(o)$ record how finding o was handled. Coverage closure requires

$$\forall o \in \mathcal{O}_{\mathrm{mat}}, \quad d(o) \in \{\text{Represented}, \text{Explained}, \text{Review}\}. \tag{4}$$

Represented links a finding to a checked episode or key moment. *Explained* records evidence that it is already covered or duplicates an existing observation. *Review* records the time and question requiring human attention. These are outcomes for findings, not additional episode types. Time overlap alone cannot explain an otherwise unrecorded operation. The requirement covers findings exposed by the scan; finite viewing can still miss events.

A missed-navigation finding. In a recorded case at 25–26 seconds, the coverage call reports departure from a stable work area, passage through a visible doorway, and arrival in another area. It creates a question about the missed movement and requests navigation context. The candidate then enters the normal visual investigation before it can change the timeline. If the relevant geometry is unclear or unseen, the call instead requests inspection of a possible change of area, without assuming that navigation occurred.

Check both changes when an activity resumes. Let P be an activity interrupted by Q and then resumed, giving $P \to Q \to P$. EEA checks both changes and whether work toward P was left unfinished. A new task requires evidence that P ended or was interrupted while Q began and continued toward a different goal. Keeping Q within P instead requires positive evidence that it helped complete P, such as a visible handoff or a state change linked to its resumption. Similar views before and after Q are insufficient. If neither interpretation is established, the boundary remains for review. A tool change, pause, or failure to prove a split does not prove a merge. Important internal operations can remain as key moments even when they do not justify another task.

Separate movement from endpoint names. Navigation requires wearer movement from one stable work area to another. A camera turn, a moving hand or object, or a step toward another area is insufficient. Claiming a particular doorway or separator was crossed additionally requires that crossing to be visible. If departure and arrival establish a change of work area but the separator is unseen, EEA keeps the supported movement without inventing the precise crossing. The room or work-area name is checked separately. For areas A and B, a supported $A \to B \to A$ route therefore retains two episodes even when the areas cannot be named or the first and last views look alike.

A.5 FROM APPROVED CLAIMS TO THE FINAL ANNOTATION

Rebuild the timeline after each accepted change. Once every claim and important coverage finding has a recorded outcome, including any review request, EEA assembles the timeline from approved details, original times, and supporting observations. Newly verified navigation or a task boundary can split an earlier episode. A description kept in either resulting piece must have support within that piece; evidence elsewhere in the original longer interval is insufficient. A point boundary creates a cut, never a zero-duration episode. Unresolvable source or partition errors stop the run rather than produce an overlapping timeline.

Write from accepted information, not from the initial story. The readout call turns approved records into descriptions without changing their meaning. It cannot add an event, precise object type, work-area name, completion state, or key moment that was not approved. Times, unknown values, broader descriptions, and local review requests are preserved. If a sentence would add meaning, the call uses the approved value directly or leaves that optional text empty. An invalid readout falls back to a fixed rendering of the approved records. Each final episode records its type, original start and end times, action and goal, relevant objects and work areas, key moments, evidence links, and any unknown values or review requests. These entries must satisfy the time and evidence checks below.

Check the final wording without adding new evidence. The final-check call receives the candidate annotation, accepted and removed claim records, evidence references, episode intervals, and review requests. It receives no video and cannot discover an event or approve a more specific interpretation. It checks that every statement agrees with an accepted record and that removed details have not returned elsewhere in the annotation. A content inconsistency causes the affected claim to be checked again; invalid source records, a broken timeline, or reintroduced rejected content block the output. A pass is a recommendation; the controller still applies the final checks below before release. A request for further checking can concern an existing claim, but cannot introduce newly discovered activity.

Any final cleanup may reorganize approved information or remove or broaden wording within the scope of a recorded issue. It cannot add visual evidence, restore a removed exact detail, or find events when coverage is disabled. Changed claims must pass the evidence and output checks again. A failed cleanup is discarded and the last validated annotation is retained; if none exists, the run fails.

Evaluate only the validated result. An annotation can be released only when no claim remains open and every important coverage finding has a recorded outcome. Human-review requests count as handled questions, but do not permit unsupported answers. The final check also requires valid source references and times, positive episode durations, complete coverage without overlap, key moments inside their episodes, and agreement between every description and its approved claim. After these checks pass, the system fixes one final version for evaluation. Evaluation uses this version alone, not the initial annotation, a discarded candidate, an earlier version, or the last model response merely because it was generated last.

Keep ablation settings effective through the final step. Disabling a component changes the complete annotation process, not just one prompt. Without verification, the targeted investigation loop is skipped, while source, response-format, and time checks remain. Without coverage, neither the coverage call nor final processing can find new events; they may still correct contradictions in already proposed claims. NoBroad retains valid exact decisions but sends broad-only information to review or unknown; final processing cannot recreate an accepted broad value. LabelVisible changes only whether the current candidates are shown during observation. Uniform replaces targeted views with equally many uniform views of the same resolution within the authorized ranges, keeping the relation question, number of actions, controller, and readout unchanged.

The four verification/coverage conditions start from the same initial observations and then run independently; none inherits Full EEA's later evidence. The three component comparisons instead retain the fixed investigation locations and order. Each condition uses its own observations and state, with no future pixels revealed early. If a range cannot provide enough unused uniform views, that step is excluded from local paired statistics and retains the original Full step in the source-level result; the replacement coverage is reported.

A.6 Model and video settings

All generation calls in the main experiment use Qwen3.8-27B. The second-backbone experiment replaces it throughout the same role sequence with Qwen3-VL-30B-A3B-Instruct. Both use BF16 weights and key/value caches without quantization, with the matching processor for each model. Their context limits are 131,072 and 65,536 tokens, respectively. Ordinary generation is non-thinking, with temperature 0.7, top-p 0.8, top-k 20, and repetition penalty 1.0. Presence penalty is 1.5 for the main model and 0 for the second. The output limit is 8,192 tokens, increased to 16,384 for timeline generation, integration, and readout. The main seed is 20260916; repeat runs additionally use 20260917 and 20260918.

The input is the full field of view from one designated monocular RGB stream, with original timestamps and time base. Images have a maximum side of 768 pixels and JPEG quality 95; a call receives at most 128 images. The routine 1-fps cache does not limit the available evidence: targeted requests can retrieve native frames, and initial observation uses 2 fps as specified in Appendix A.3. Generation receives no audio, narration, gaze, depth, trajectory, external view, reference annotation, video-specific task name, supplied question, or event list.

A.7 Original prompts used by EEA

The following nine prompts are reproduced with their complete instruction text unchanged. They follow the order of the annotation process. Initial construction and later revision share one prompt, whose two call modes are defined within the instructions. Each prompt is accompanied during

execution by the views or records allowed in Table 5, the remaining work allowance where applicable, and the structured response described in Table 5. Thus, the prompt supplies the instructions, while the current video and investigation supply the changing input.

In these instructions, a *slot* is one annotation detail, *lineage* is its history of claims and supporting observations, and *residue* means observations not yet linked to the timeline. The fixed *ontology* is the set of broader-description links in Table 6; the *canonical* annotation is the single final timeline. *Gold* means the human reference annotation, which is never supplied to the generation calls.

A.7.1 ORIGINAL PROMPT FOR INITIAL VIDEO OBSERVATION

This call records visible facts and their times before any task interpretation is proposed.

```
You are the Literal Observer for an untrimmed egocentric video. Use only
    the source-bound frames or clips supplied in the request. Report what
     is visibly present and when it is visible. Do not infer a task goal,
     intent, completion, tool function, room name, navigation, or object
    identity across an unseen gap.

Record concise, atomic observations of:
- hand--object and object--surface contact;
- relative motion and visible use;
- object or display state before and after an action;
- wearer/camera movement, stable work areas, and visible thresholds;
- occlusion, darkness, missing phases, and other visibility limits.

One observation should describe one temporally coherent operation. Split
    ordered operations such as open-then-extract or pick-up-then-cut. Do
    not split one continuous instance merely because the view changes.
    Use broad physical nouns unless diagnostic structure supports a
    subtype.

A changed background does not by itself establish navigation. Navigation-
    related observations must distinguish wearer movement from object/
    hand motion and must report departure, threshold or separator,
    movement, and stable arrival only when those relations are visible.
    An A->B->A path contains two ordered movements.

Observation IDs are content-free ordinals. Every fact must cite the
    supplied evidence reference and source timestamp/anchor available in
    the request. If a relation is not visible, state the missing relation;
     never fill it from scene knowledge, filenames, prior annotations,
    candidate labels, Gold, or expected answers.

Return JSON only and conform exactly to the supplied response schema. Do
    not expose chain-of-thought.
```

A.7.2 ORIGINAL PROMPT FOR BUILDING AND REVISING CLAIMS

The same instructions support initial construction and later revision; the call mode determines which part applies.

```
You build or revise tentative annotation claims from source-bound literal
     observations. A claim is one proposed public detail over one native-
    time interval, such as an action, object role, state, task boundary,
    navigation occurrence, work-area identity, or key moment.

CALL_MODE=INITIAL:
- group observations into tentative Task and Navigation episodes;
- preserve important internal operations as timestamped key moments
    unless evidence supports a task split;
- propose the most specific description justified by the cited
    observations;
- list assumptions and materially different alternatives separately;
- mark unsupported, unseen, or unresolved details as uncertain rather
    than completing them from context.
```

```
CALL_MODE=REVISION:
- use the new blind observation and existing claim lineage;
- revise only the selected public slot and interval;
- keep supported facts unchanged;
- either propose a revised claim, request one new discriminating relation,
     or stop with a broad/unknown/review recommendation;
- never treat the blind observer's wording, model agreement, or
    confidence as evidence beyond its cited source pixels.

A task groups actions serving one coherent goal. Tool change, pause,
    camera turn, or short carry does not by itself create a new task.
    Navigation requires wearer movement between stable work areas; object
     motion or local repositioning is not navigation.

Do not use Gold, reference annotations, expected outputs, filenames,
    directory task names, or prior human answers. Return JSON only under
    the supplied schema. Do not expose chain-of-thought.
```

A.7.3 Original prompt for checking a proposed claim

This call asks which alternative interpretation and visual relation could change the selected claim.

```
You challenge exactly the public claim and slot selected by the
    controller. Treat the claim as replaceable. Identify:
1. the unsupported assumption in the selected claim;
2. at least one materially different alternative that would change the
    annotation;
3. the visible relation that would distinguish the claim from that
    alternative;
4. pixels that could refute the selected claim or force a safer broader
    description.

Stay within the controller-issued slot and interval. Do not open
    unrelated questions about purpose, completion, object identity, task
    boundary, navigation, or destination unless that dimension is
    explicitly in scope. Paraphrases that preserve the same structured
    fact are not conflicts.

Check especially for holding mistaken for use, presence mistaken for a
    target role, identity carried across an unseen gap, completion
    inferred from a final view, unsupported room/object subtype, missed
    return navigation, false task split/merge, and confident prose
    spanning dark or occluded time.

Formulate neutral visual questions. Do not ask whether the proposed label
     is correct. Ask what contact, relative motion, state change,
    continuity, boundary, or navigation geometry the video must show. Do
    not replace one hidden purpose with another plausible story.

Return JSON only under the supplied schema. Do not expose chain-of-
    thought.
```

A.7.4 Original prompt for selecting the next view

The planner chooses one new view that can answer the current visual question.

```
Choose exactly one controller-authorized visual action that can answer
    the current discriminating relation. The request must expose relevant
     source pixels not already inspected for this conflict.

Select the smallest sufficient view:
- source frame or crop for static structure/contact;
- short continuous clip for relative motion, contact onset/release, or
    local state change;
- wider before--action--after context for completion;
- gap-continuity or lifecycle view for identity across time;
```

```
- boundary context for task split/merge;
- navigation context for departure, threshold, movement, arrival, and a
    possible return leg.

Respect the allowed source interval, native timestamps, media identity,
    remaining action budget, and tool capabilities. Do not choose a
    repeated view, an unauthorized time range, or a view that could
    change only an out-of-scope narrative detail. A tool request
    retrieves pixels; it does not decide the semantic answer.

Explain the expected decision effect in one concise sentence. Return JSON
     only under the supplied schema. Do not expose chain-of-thought.
```

A.7.5 ORIGINAL PROMPT FOR OBSERVING WITHOUT CANDIDATE LABELS

The model receives the new views and relation question, but not the proposed answer.

```
Inspect only the newly opened source-bound views and the neutral relation
     question. You must not receive or infer the candidate labels,
    initial annotation, proposed answer, losing hypothesis prose, Gold,
    reference answer, filename, or expected output.

First report atomic literal observations with evidence references. Then
    report the requested relation as:
- PRESENT: the relation is visibly established;
- ABSENT: the shown scope is sufficient to rule the relation out;
- AMBIGUOUS: the relevant action/region is shown but the relation cannot
    be decided;
- NOT_VISIBLE: the time, phase, or region needed to decide is not shown.

AMBIGUOUS and NOT_VISIBLE are not evidence of absence. Do not upgrade
    holding to use, a final state to completion, similar-looking objects
    to one identity, camera motion to navigation, or scene geometry to a
    room name. For navigation, distinguish wearer/camera translation from
     object/hand motion and report the visible departure, separator,
    passage, and arrival separately.

Search the supplied view for counterevidence to the requested relation
    and state whether competing alternatives were actually excluded. Do
    not output a final annotation label or release decision.

Return JSON only under the supplied schema. Do not expose chain-of-
    thought.
```

A.7.6 ORIGINAL PROMPT FOR RECOMMENDING A DECISION

This call recommends what to retain or review; the controller makes the acceptance decision.

```
You propose one terminal disposition for the selected claim using only
    the normalized candidate claims, source-bound blind observations,
    evidence capabilities, and controller-supplied allowed resolutions.
    The deterministic controller will validate your proposal against the
    frozen evidence policy.

Use these meanings:
- ACCEPT_EXACT: the exact claim has positive visual support, required
    counterevidence was checked, and the live alternatives were excluded;
- ACCEPT_BROAD: a frozen ontology supplies one common public-safe
    description, that description has its own positive visual support,
    and remaining uncertainty does not change the usable annotation;
- BROAD_WITH_HUMAN_REVIEW: the same supported broad description can be
    retained, but an unresolved distinction still changes an important
    detail;
- CANNOT_IDENTIFY_WITH_HUMAN_REVIEW: no supported description settles an
    important detail or the activity itself;
```

```
- CLEAR/REJECT: evidence contradicts the proposed public detail; remove
    that detail without deleting other supported episode information.

Never accept an exact value because it was not disproved. Never invent a
    broad parent, choose between tied parents, use model confidence, or
    let a review mark keep an unsupported exact guess. Every retained
    fact must cite its supporting evidence. If a permitted terminal
    decision cannot be justified, choose the safest allowed review/
    unknown outcome rather than a likely answer.

Return JSON only under the supplied schema. Do not expose chain-of-
    thought.
```

A.7.7 ORIGINAL PROMPT FOR FINDING MISSED ACTIVITY

The coverage call searches source observations and coverage records, without reading the annotation narrative.

```
Review the source-spanning literal observations, residue ledger, timeline
     coverage map, and observation mappings to find material activity
    that is not yet represented. You may see which observations and
    intervals are covered, but not the annotation's prose, candidate
    labels, Gold, reference answers, or expected output.

Look for:
- a distinct operation missing inside an existing Task;
- a sustained change of goal that may require a task boundary;
- wearer movement between work areas, including an outward and return leg
    ;
- dark, occluded, or weakly covered intervals hidden by confident
    neighboring text;
- an important observation with no link to an episode, key moment, or
    documented duplicate explanation.

Emit one obligation per material public decision. Do not combine
    different operations or combine a task activity with a navigation
    crossing. A new finding is not an accepted fact: specify the literal
    basis, missing relation, interval, and recommended visual tool so
    that it can re-enter the normal investigation.

Navigation is MISSED_NAVIGATION only when the required geometry is
    visibly present. If the relevant phase is unclear or unseen, emit an
    unresolved-change obligation and request navigation context; do not
    infer a crossing. Same-area repositioning and object motion are not
    navigation.

Return all material observations as either already mapped or as explicit
    obligations. This is closure over observations exposed by the scan,
    not a guarantee that finite sampling found every event. Return JSON
    only under the supplied schema. Do not expose chain-of-thought.
```

A.7.8 ORIGINAL PROMPT FOR WRITING THE FINAL ANNOTATION

This call expresses approved records as the final annotation without adding facts.

```
You receive only controller-approved structured episode records, native
    timestamps, evidence links, and localized review fields. Render them
    into the public annotation schema.

Do not read or reuse the initial annotation, losing claims, rejected
    details, candidate labels, model confidence, Gold, reference answers,
     or any visual fact not present in the approved records. Do not add a
     new event, boundary, object subtype, room name, intent, completion
    state, navigation endpoint, or key moment. Keep broad values broad.
    Preserve unknown and human-review fields exactly. Use half-open
```

```
    episode intervals and the supplied source timestamps without moving
    boundaries.

If a natural-language field cannot be rendered without adding meaning,
    use the approved structured value verbatim or leave the optional text
     field null. Return JSON only under the supplied public schema. Do
    not expose chain-of-thought.
```

A.7.9 ORIGINAL PROMPT FOR CHECKING THE FINAL TEXT

This call checks consistency with approved records; it neither inspects video nor authorizes output.

```
Perform a text-only consistency audit of the single canonical public
    annotation. You receive terminal claim lineage, evidence references,
    episode intervals, review obligations, and the public projection. You
     receive no images, video, Gold, reference annotation, candidate
    labels, losing hypotheses, or prior draft prose.

Check only what the supplied records can establish:
- every public fact has one terminal accepted lineage;
- wording is no more specific than the approved structured value;
- removed or losing claims do not reappear in any public field;
- action, role, completion, room, boundary, navigation, and key-moment
    wording agrees with the corresponding terminal slots;
- timestamps and evidence references match the claim lineage;
- episodes form one positive-duration, gap-free, nonoverlapping timeline;
- review/cannot-identify obligations are represented and not hidden by
    confident prose;
- all coverage obligations are mapped, explained, or terminalized for
    review.

You cannot inspect pixels, discover a new event, create a new boundary/
    navigation claim, rewrite the annotation, or authorize release. If an
     existing claim's wording or lineage is inconsistent, request
    REOPEN_CONFLICT for that claim. If the timeline, provenance, or
    dropped-claim leakage is invalid, request BLOCK_RELEASE. Otherwise
    return PASS/fixed-point candidate. Release is decided by the
    deterministic validator.

Return JSON only under the supplied schema. Do not expose chain-of-
    thought.
```

A.8 ORIGINAL PROMPTS FOR BASELINE ANNOTATION AND ADAPTATION

The prompts below implement Direct, Dense-Group, Dense-Group+SR, the task adaptations for DeMiAn-Ann and VideoHV-Ann, and their common readout. They use each baseline's own observations or answers, not EEA's accepted claims. The common output format allows all methods to retain navigation, unknown values, and local review; it does not apply EEA's evidence controller to the baselines.

Direct observes nonoverlapping 120-second windows at 0.5 fps and then combines the window annotations. Dense-Group uses the development-selected 120-second windows at 1 fps with 25% overlap, first producing detailed observations and then grouping them. Dense-Group+SR critiques and revises that annotation for two rounds using only frames already seen by Dense-Group. These steps appear below in execution order.

DeMiAn-Ann first forms its own context in 60-second windows at 0.5 fps and automatically proposes continuous action segments. Each segment supplies ten uniformly sampled native frames to the original four DeMiAn prompts covering the current action, environment, arm state, and brief reasoning. VideoHV-Ann forms annotation questions and five candidate answers from its own context, then retains the original sequence of hypothesis construction, distinctness checking, regeneration, verification, and selection. It uses a full-video pool of 180 uniformly sampled native frames, at most three rounds, a distinctness threshold of 0.5, and two verifier tool iterations per round. The two task-adaptation prompts reproduced below specify how those original procedures are used and how

their results become timeline annotations; they are not presented as replacements for the original methods' internal prompts.

A.8.1 DIRECT: ORIGINAL PROMPT FOR EACH VIDEO WINDOW

This prompt annotates a uniformly sampled window using its original video times.

```
You receive one nonoverlapping window from an untrimmed egocentric video,
     sampled uniformly at 0.5 fps, together with its original source
    timestamps. Describe every Task and Navigation episode visible in
    this window using global source time. Group actions serving one goal;
     preserve movement between work areas as Navigation; use unknown or
    localized review when the video does not support a detail. Do not use
     filenames, task-directory names, Gold, reference boundaries,
    external questions, or events outside the shown window.

Return JSON with window_start, window_end, episodes, key_moments,
    unknowns, and review_intervals. Do not treat the window edges as task
     boundaries unless the shown activity supports a boundary.
```

A.8.2 DIRECT: ORIGINAL PROMPT FOR COMBINING WINDOW ANNOTATIONS

The integration call combines only the window outputs, without inspecting new video.

```
You receive only the ordered outputs of nonoverlapping video windows.
    Produce one global Task--Navigation timeline in the original source
    clock. Resolve duplicate descriptions near adjacent windows, preserve
     explicit Navigation and return movements, and keep unknown/review
    content. Do not infer an event absent from the window outputs,
    request new video, or turn a window edge into an episode boundary by
    default.

Return the common public annotation schema only.
```

A.8.3 DENSE-GROUP: ORIGINAL PROMPT FOR DETAILED WINDOW OBSERVATIONS

This prompt records actions and movement, including brief operations near window edges.

```
You receive one overlapping window of an untrimmed egocentric video at
    the configured sampling rate and original timestamps. Produce a dense,
     time-ordered account of visible actions, object interactions, state
    changes, wearer movement, stable work areas, and uncertainty.
    Preserve brief operations and returns near window boundaries. Do not
    infer a task boundary merely because the window ends.

Return JSON with atomic observations, tentative task cues, navigation
    cues, key-moment candidates, and visibility limits.
```

A.8.4 DENSE-GROUP: ORIGINAL PROMPT FOR GROUPING OBSERVATIONS

The grouping call turns the window observations into tasks and navigation, without treating window edges as task boundaries.

```
You receive all dense window outputs in source-time order. Group the
    visible actions into coherent Task episodes and Navigation episodes
    covering the full video. Merge overlap duplicates; do not merge
    operations serving different goals; do not split only because a
    window boundary, tool change, short pause, or camera turn occurs.
    Preserve important internal operations as key moments. Use unknown/
    review when the dense descriptions do not support a detail.

Use only the supplied window outputs; do not inspect new video. Return
    the common public annotation schema.
```

A.8.5 DENSE-GROUP+SR: ORIGINAL PROMPT FOR CRITIQUE

The critique call receives the Dense-Group annotation and exactly the frames used to produce it.

```
You receive the selected Dense-Group draft and exactly the same frames
    previously shown to Dense-Group. Identify unsupported specificity,
    inconsistent boundaries, omitted content visible in those frames,
    duplicated episodes, missing Navigation, and contradictions between
    text and pixels. You may not request or receive any new frame,
    enlarge a time interval, or use Gold/reference information. State
    each criticism with the affected episode/field and cited existing
    frame IDs.

Return JSON only.
```

A.8.6 DENSE-GROUP+SR: ORIGINAL PROMPT FOR REVISION

Revision uses the critique and the same observed frames; it cannot acquire new visual evidence.

```
Revise the Dense-Group draft using only the supplied critique and the
    same original frame set. Correct supported errors, preserve supported
     content, and use unknown/review when the existing pixels cannot
    settle a criticism. Do not add facts that require unseen pixels. Keep
     source timestamps and produce the common public annotation schema.
    This prompt may be run for the fixed number of refinement rounds
    selected on development data.
```

A.8.7 DEMIAN-ANN: ORIGINAL TASK-ADAPTATION PROMPT

This is the adaptation instruction used around the original DeMiAn procedure, not a replacement for its four aspect prompts.

```
This adapter does not replace the upstream DeMiAn prompts. Use the frozen
     upstream commit and its original prompts for exactly four aspects:
    current_action, environment-simple, agent_arm_state, and reasoning-
    short. The upstream prompt receives the method's own 60-second/0.5-
    fps context and ten native uniformly sampled frames for each
    automatically generated continuous action segment.

After the four aspect outputs are produced, map only those outputs and
    their segment timestamps to the common Task--Navigation schema. Do
    not inspect EEA outputs, Gold, reference boundaries, or additional
    video. Do not invent an event that is absent from the DeMiAn results.
     Window or segment edges do not automatically become Task boundaries.
```

A.8.8 VIDEOHV-ANN: ORIGINAL PROMPT FOR GENERATING QUESTIONS

This prompt produces an annotation-relevant question and five candidate answers from the method's own context.

```
Given the method's own context for one video window, generate one
    annotation-relevant question and exactly five materially different
    candidate answers. Questions should concern the visible action, task
    change, completion, object role, or Navigation needed for a Task--
    Navigation timeline. Do not use Gold, reference annotations, supplied
     external questions, or EEA outputs. A suggested time range guides
    the question but does not restrict the verifier's authorized full-
    source candidate-frame pool.

Return JSON with question, five candidates, suggested interval, and
    annotation slot.
```

A.8.9 VIDEOHV-ANN: ORIGINAL TASK-ADAPTATION PROMPT

The adaptation retains the hypothesis-verification procedure and converts selected answers into the common annotation format.

```
Run the frozen upstream VideoHV pipeline at its fixed commit: hypothesis
    construction, distinctness checking, candidate regeneration when
    distinctness is below 0.5, tool-based verification, and final
    selection. Use at most three rounds and two verifier tool iterations
    per round. The verifier may select from the fixed pool of 180
    uniformly sampled native frames covering the full video. Use the
    common backbone for captioning and all generative roles.

After selection, map only selected answers and their source-time context
    to the common Task--Navigation schema. Do not expose losing
    candidates, EEA controller decisions, Gold, or reference annotations
    to the public readout. The readout may use unknown/review but may not
     request additional video.
```

A.8.10 ORIGINAL PROMPT FOR THE SHARED BASELINE READOUT

This prompt expresses only the selected baseline's own outputs and does not apply EEA's controller.

```
Convert only the supplied method-specific outputs into the common public
    Task--Navigation annotation schema. Preserve the method's own
    timestamps, unknown values, review decisions, Navigation, and key
    moments. Do not inspect video unless the baseline definition
    explicitly provides it to this readout. Do not add events or facts
    from other methods, Gold, reference annotations, filenames, or common
     sense. Do not silently repair a method by applying EEA's evidence
    controller.

Return JSON only under the common public schema.
```

B EXPERIMENTAL PROTOCOL

We follow the main experiments through videos, comparisons, scoring, quality, mechanisms, and human work, then test another backbone, seeds, and temporal thresholds.

B.1 EGOEASY-BENCH: COMPLETE ANNOTATION FROM AN UNTRIMMED EGO STREAM

What the benchmark asks a system to produce. EgoEasy-Bench (EEB) requires a complete task–navigation timeline from one original first-person RGB file, without supplied questions or event lists. Systems must find episodes, describe visible actions, object/tool roles and work areas, locate important internal operations, and retain uncertainty where video cannot establish a distinction. References include Task, Navigation, Idle, and CannotIdentify. We score Task/Navigation recovery; Idle and CannotIdentify preserve background and unclear activity without inventing actions.

Collection scale and annotation detail. Table 9 summarizes all valid EEB data: 472 of 565 delivered video–reference pairs, 38.151 hours, and 3,620,614 native frames across all splits. Their activities include hotel room service, bathroom/floor cleaning, kitchen preparation, washing, organization, and movement through corridors, elevators, and indoor/outdoor work areas. These examples come from human annotations; they do not define a fixed action vocabulary or verify a number of independent sites. The annotations describe goals and functional roles; directory names do not become action labels or model inputs.

Frame counts are checked against media headers and reference metadata, not treated as independent samples or model inputs. All valid sources have 1920×1200 resolution. Annotations use native presentation timestamps with a $1/90{,}000$-second time base; this clock resolution does not imply equivalent human boundary precision. Total durations from reference endpoints and media streams differ by 1.398 seconds; both round to 38.151 hours. Scoring uses original timestamps, not nominal frame rate. Collection groups and parent recordings do not establish independent people or continuous sessions. Appendix B.2 gives the 428-file test split and its references.

Why a plausible summary is insufficient. The benchmark combines different temporal requirements in the same input. Although most files last about five minutes, over half of navigation episodes last

Table 9: EEB's full valid collection, including development and test. These are not the 428-file test denominators. Raw key moments are representative timestamps before event-group compilation; evidence windows are supporting intervals, not additional events.

Property	Full valid collection
Original files / duration	472 / 38.151 hours
Native source frames / resolution	3,620,614 / 1920 × 1200
Collection-directory groups / parent recordings	15 / 41
Median file duration	300.117 seconds
Task / Navigation episodes	1,756 / 992
Idle / CannotIdentify episodes	106 / 33
Raw key moments / evidence windows	4,384 / 4,723
Files with multiple tasks	321/472 (68.01%)
Files containing navigation	258/472 (54.66%)
Median task / navigation duration	26.205 / 4.522 seconds
Navigation episodes lasting at most five seconds	534/992 (53.83%)

at most five seconds, and 177/992 (17.84%) last at most two seconds. Describing the surrounding activity is therefore insufficient: brief departures and returns must each be recorded. Conversely, 68.01% of files contain multiple tasks, so a single broad activity label cannot replace finding changes of objective. Preparation, repeated movements, checks, and cleanup must still be grouped when they serve one objective, rather than split at every action or sampling window.

The descriptions must be checked as well as the times. Carrying a tool does not establish using it; opening a door does not establish walking through it; returning to a familiar view still requires a return-navigation episode. References retain action, tool/target role, navigation origin/destination, carried objects, and important internal operations. When occlusion hides a relevant detail, the annotation may leave it unknown rather than guess. The valid collection includes 33 CannotIdentify episodes and two Task episodes marked for human handling; the presence of an uncertain detail does not make the whole video unusable.

Evidence supports the reference without becoming an inference shortcut. The 4,723 evidence windows link episodes to source intervals, frame ranges, optional crops, and the details they support. Together with human key moments, they show which video evidence underlies the reference. They are never supplied to the generation methods. A prediction inside an evidence window does not automatically count as a recovered event: it must name the same operation and either match a human timestamp or have confirming frames near its own timestamp (Appendix C.2). EEB does not provide exhaustive structured state/completion obligations; state-change events and unsupported completion claims are still checked. Its scope is joint task/navigation structure, necessary semantics, event coverage, and appropriate uncertainty, not an exhaustive annotation of every visible attribute.

One file illustrates the combined requirement. Table 10 summarizes a 299.937-second human reference from the valid collection. The wearer organizes a service trolley, enters a bathroom for different cleaning tasks, returns, and makes further trips carrying towels and waste. This is a human reference, not a prediction.

Calling this file "hotel cleaning" would miss goal changes and repeated trips; treating every hand movement as a task would lose coherent objectives. Both task grouping and distinct navigation episodes are required.

Relation to established egocentric benchmarks. EPIC-KITCHENS-100 provides narrations and temporally localized actions (Damen et al., 2022). Ego4D Goal-Step already connects goals, steps, and substeps and includes completion and relevance annotations (Song et al., 2023). HD-EPIC supplies detailed, interconnected recipe, action, object, and multimodal references (Perrett et al., 2025); Ego-Exo4D adds synchronized first- and third-person views and skilled-activity annotations (Grauman et al., 2024). Thus, neither temporal localization nor hierarchical activity understanding is unique to EEB.

EEB tests these decisions together from one egocentric video without supplied questions: supported descriptions, missed operations, goal-based boundaries, each navigation episode, and justified detail. Accurate descriptions of selected moments or a complete-looking account with unsupported claims are insufficient. EEB tests the combined task, not greater size or duration. HD-EPIC adds longer recordings under the same protocol; these statistics do not establish which benchmark is harder.

Table 10: One EEB human reference from the full valid collection, not a model result. Times are rounded for display; evaluation retains native timestamps. Each row summarizes an original episode rather than creating new scoring labels.

Time (s)	Type	Reference activity
0.000–28.641	Idle	Remain beside the room-service trolley.
28.641–112.517	Task	Organize cleaning tools, cloths, and supplies on the trolley.
112.517–116.807	Navigation	Move from the corridor trolley area into the bathroom.
116.807–132.306	Task	Wipe the toilet with a cloth.
132.306–175.439	Task	Clean and organize the vanity, sink, and mirror.
175.439–182.871	Navigation	Return from the bathroom to the corridor trolley.
182.871–229.219	Task	Fold and stack towels.
229.219–231.461	Navigation	Carry towels from the trolley into the bathroom.
231.461–240.466	Task	Arrange the towels beside the bathtub.
240.466–249.453	Navigation	Move from the bathtub work area to the guest-room desk area.
249.453–266.259	Task	Collect waste, arrange the bin liner, and gather linen.
266.259–273.718	Navigation	Carry the bin and linen back to the corridor trolley.
273.718–299.937	Task	Empty the bin and organize linen and bags.

B.2 Test splits, common references, and development isolation

One original file is one inference input. Inputs are complete official HD-EPIC RGB recordings and original EEB files (Appendix B.1), never cut at human event boundaries. Related EEB files remain correlated and are not concatenated without verified continuity. Table 11 separates test, development, and reserve files. Test duration is 51.817 hours; EEB collection-wide statistics also include 44 valid non-test files.

Table 11: Data allocation. Hours are rounded after summing frozen media durations. Collection-directory groups are dependence proxies, not verified person or site counts.

Benchmark/use	Files	Hours	Highest-level groups
HD-EPIC development	8	1.119	2 wearers
HD-EPIC test	60	17.010	7 wearers
EgoEasy evaluation development	8	0.633	3 directory groups
EgoEasy development reserve	36	2.712	3 directory groups
EgoEasy test	428	34.806	11 directory groups

HD-EPIC uses P01–P02 for development and P03–P09 for testing. EgoEasy test files span 32 parent recordings. Evaluation development and reserve overlap in two collection-directory groups and seven parent directories; neither overlaps test at either level. Model inputs use identifiers that reveal no directory names, task labels, or annotation status. Separate directories do not prove separate people or physical sites.

Screening precedes the compared predictions. The 93 excluded EEB files are marked invalid in the original delivery. Reasons include inadequate visible behavior or source pixels, severe occlusion, and repetitive or insufficiently structured activities. Screening does not depend on method scores; it defines a task-focused collection rather than a random sample of everyday video. The 472 valid files divide into 8 evaluation-development, 36 reserve, and 428 test files. The collection-wide counts in Table 9 must not replace the test denominators in Table 12.

Human references define the annotation target. The 120-person annotation pool is organized into three cross-checking groups. Both benchmarks' references cover coherent task/navigation episodes, important internal operations and changes, and necessary visually supported core facts. Reference makers do not see method predictions. HD-EPIC's official activities, recipes, and narrations assist human omission checks; they neither automatically define coherent tasks nor enter model generation. This is not three-fold validation or 120 independent judgments per item.

Preparation, repetition, checking, cleanup, short pauses, and local movement can remain within one task objective. Navigation requires wearer movement between stable work contexts; a turn of the camera, movement of an object, or opening a door without crossing does not establish it. A return is a separate occurrence even when place names are unknown. Core distinctions include carrying versus using, opening versus crossing, and starting versus completing. Decorative color, material,

ordinary subtype, and appearance modifiers are not mandatory facts. Truly unobservable distinctions may remain unknown.

A fixed model call extracts required facts and operations from these human annotations (Appendix C.1). Several timestamps marking the same operation can form one event group; every original key moment belongs to a group. Table 12 reports the counts used for testing.

Table 12: Frozen test references. Exact obligations can coincide with minimum useful meaning; only the independently finer subset demands an additional core distinction.

Reference item	HD-EPIC	EgoEasy
Task episodes	624	1617
Navigation episodes	238	922
Necessary core fields	3184	6742
Raw key moments	2240	3953
Core event groups	1768	3800
Observable exact-field obligations	2198	5100
Obligations with independently finer core meaning	466	53
Premarked round trips	76	240
Observable navigation endpoints	428	1806
Premarked motion-negative intervals	196	486

Evaluation development does not tune EEA. EEA was not fine-tuned or behaviorally adjusted using these videos. Evaluation development checks the shared scorer; baseline configuration selection is a separate, limited procedure. Earlier exploratory outputs were inspected and remain separately recorded, including the original four-source Direct outputs and failures. This is not a claim that no test output was ever viewed. Dense-Group and refinement settings are selected only on the original eight EgoEasy development files; EEA's behavior is not selected using test feedback.

The mechanism, second-backbone, and enhanced-resource experiments share 60 preselected files, 30 per benchmark, covering all 7+11 groups and 8.909/2.401 hours. The human study uses 36 files, 18 per benchmark, covering 7+11 groups and totaling 7.319 hours. The seed study uses 12 files, six per benchmark, covering 5+6 groups. Selection uses group/parent identities, duration ranks, and fixed hashes before test predictions; each EgoEasy subset contains at most one file per parent recording. Existing Full outputs are reused rather than regenerated to choose a favorable subset result.

B.3 Common model access and executable comparisons

Hold the model and access permissions fixed. Table 13 lists inference settings; Table 14 fixes model and upstream versions. All main generation roles—captioning, drafting, planning, observation, verification, repair, and readout—use Qwen3.8-27B. The second-backbone study changes every role to Qwen3-VL-30B-A3B-Instruct. Processors and weights share a snapshot; weights and key/value caches use BF16 without quantization.

Methods receive only the designated ego RGB stream, native timestamps, and common task/output definitions. No gold, reference boundaries, supplied questions/events, task-directory names, narration, audio/ASR, gaze, depth, trajectories, or external views enter generation. All methods may emit navigation, unknown values, and localized review. Equal video access permits different frame choices; identical evidence is required only in the specified component controls and self-refinement comparison.

Table 13: Fixed inference configuration. Upstream role-specific sampling settings are preserved where explicitly specified and recorded per call.

Item	Setting
Image transport	Full designated monocular view; maximum side 768; JPEG quality 95.
Frame access	Routine 1-fps cache plus on-demand native frames; at most 128 images per call. Native timestamps determine time.
Ordinary generation	Non-thinking; temperature 0.7, top-p 0.8, top-k 20, repetition penalty 1.0; presence penalty 1.5/0 for main/second model.

Table 13 (continued)

Item	Setting
Output limits	8,192 tokens ordinarily; 16,384 for timeline generation/integration/readout; 256 per DeMiAn aspect.
Seeds	20260916; repetitions additionally use 20260917 and 20260918.
Software	vLLM 0.19.1; PyTorch 2.10.0; Transformers 5.8.1; NumPy 2.2.6; SciPy 1.17.1.
Deployment	RTX6000 Ada, 48 GB. Main: two-device tensor parallelism, 131,072 context tokens. Second: one device, approximately 24 GiB CPU expert offload, 65,536 context tokens.

Table 14: Frozen model and upstream revisions. Model repository names use the Qwen organization.

Model / upstream method	Weight revision / commit
Qwen3.8-27B	`1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`
Qwen3-VL-30B-A3B-Instruct	`9c4b90e1e4ba969fd3b5378b57d966d725f1b86c`
DeMiAn	`2a823692e63bb4eb4c44831e427956a7ba90fdfb`
VideoHV	`ddc160bd05e2d64922ca717d1834500fa7d8f492`

Direct observation, dense grouping, and refinement. Direct processes nonoverlapping 120-second windows at 0.5 fps and integrates their outputs without further video. Dense-Group produces descriptions of overlapping windows and groups them into coherent episodes; window edges do not automatically become task boundaries. Its development grid is 30/60/120 seconds $\times$ 0.5/1 fps with 25% overlap. Dense-Group+SR critiques the selected draft using its original frame hashes and performs one or two refinement rounds; it receives no new frames.

For each configuration, we combine episode segmentation F1 (Seg) and necessary-fact F1 lower bound (Field_L) into the selection score

$$J = \frac{2\,\mathrm{Seg}\,\mathrm{Field}_L}{\mathrm{Seg} + \mathrm{Field}_L}, \tag{5}$$

We maximize J on development videos; a zero denominator gives $J = 0$. Ties within 10^{-12} favor lower UFR upper bound, fewer tokens, then fixed configuration order. This selects 120-second/1-fps Dense-Group windows and two refinement rounds. Among five baselines, the same rule selects Dense-Group for human, resource, and seed studies ($J = 37.25\%$). The 36 reserve files are excluded.

Adapting neighboring methods without supplying their missing inputs. DeMiAn-Ann first generates its own 60-second/0.5-fps context and continuous action segments, samples ten native frames per segment, and applies the original `current_action`, `environment-simple`, `agent_arm_state`, and `reasoning-short` prompts. A restricted readout uses only this method's results.

VideoHV-Ann also generates its own context, then questions and five candidates per window. It preserves hypothesis construction, distinctness checking, regeneration, tool verification, and selection. The candidate pool contains 180 uniformly sampled native frames from the entire video; question-suggested intervals do not restrict that access. It uses at most three rounds, distinctness threshold 0.5, and two verifier tool iterations per round. Captioning and all other generative roles use the common backbone. Readout consumes only selected answers, not EEA's release controller.

The upstream commits in Table 14 fix the reused method versions. These comparisons evaluate annotation-task adaptations, not native robot-learning or question-answering scores. The frozen Open-AoE implementation audit did not locate a directly executable, complete automatic production entry point; this reproducibility finding is not a performance judgment.

EEA scales its work budget with video length, not reference answers. EEA uses the length-based scheduling and limits in Appendix A.3, including 2-fps initial observation over every window. That appendix defines the multiplier s, lists every request and runtime limit, and specifies the change scan. Assigning every frame to a window does not mean displaying every frame.

Reaching these budgets invokes the stopping rules, not weaker evidence requirements. Input transport and timestamp handling do not change the method's exact/broad/review decisions.

Keep output adaptation separate from annotation improvement. The common adapter translates representation without adding events, changing meaning, or moving boundaries. Idle in the human reference means observable background. Predictions have no Idle type, so a shared conversion treats a Task describing only observed rest/waiting as background without consulting the reference. Its saved label, wording, and times stay unchanged. Unidentifiable activity is not relabeled idle. All internal cuts remain eligible for boundary scoring.

C EVALUATION: FROM PUBLIC CLAIMS TO SCORES

The scorer separately determines what the reference requires, whether a published statement is supported, and whether it recovers the required fact at the right time. A true but unrelated statement cannot earn recovery credit.

C.1 SHARED REFERENCES AND ANONYMOUS PUBLIC ASSERTIONS

A fixed Qwen call reads only the human reference annotations and extracts the facts and distinct operations each prediction should recover, while preserving human boundaries. Repeated timestamps can be grouped only when every pair describes the same operation, role, and important outcome within one episode. Different phases, opposing states, and distinct navigation occurrences remain separate. Every raw key-moment identifier belongs to exactly one group.

A separate call extracts checkable statements from each anonymous output, without reading the references. It covers prose, action lists, objects/tools and their roles, states, places, endpoints, and key-moment text. Equivalent assertions within an episode are deduplicated; different roles, directions, or occurrences are not. Decorative exclusions and original members remain traceable, and predicted times never change.

Each core assertion receives one of four judgments: supported; contradicted; insufficient evidence for the asserted precision; or indeterminate. If the text check does not establish support, the evaluator inspects source frames. Omission from the reference or sampled frames alone does not establish error. Marking a claim for review does not exempt it from scoring.

A supported statement earns recovery credit only if it describes the required fact, role, direction, occurrence, and time. For example, a supported statement about reaching a doorway does not recover opening it. Confirmed true extras outside the required facts are neutral: neither recall credit nor false-positive penalty. Incorrect extra core assertions remain penalized. EgoEasy has no exhaustive state/completion field reference; incidental prose does not create state obligations, while important state-change events and incorrect completion claims remain scored.

C.2 MATCHING, DENOMINATORS, AND UNCERTAINTY

Common counting. A valid one-to-one match is a true positive (TP); a scored unmatched prediction is a false positive (FP), and a missed required reference item is a false negative (FN). Matching first maximizes the number of pairs, then temporal overlap or the specified tie-break. Within each benchmark, we pool counts before computing

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{6}$$

A truly empty denominator is N/A. Missing predictions score zero recovery wherever reference obligations exist; those obligations are not removed.

Episodes, cuts, and navigation. For predicted interval P and reference interval G,

$$\mathrm{tIoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}. \tag{7}$$

Here $|\cdot|$ denotes interval duration. Seg-F1 requires the same Task/Navigation type and tIoU ≥ 0.5. Background does not enter the target-episode denominator. Adjacent same-type tasks are not merged before matching. Task-only and navigation-only F1 restrict this same procedure to their respective types; navigation existence credit does not require correct place names.

Boundary-F1 matches all internal cuts one-to-one within one second, including cuts involving background or CannotIdentify, but excludes common video endpoints. For a predicted boundary at time t, distance to an allowed reference range $[a, b]$ is $\max(a - t, 0, t - b)$. The test sets have 802 HD-EPIC and 2,237 EgoEasy reference cuts.

Round-trip recall is the fraction of premarked round trips for which both legs pass navigation matching. Endpoint recovery is the fraction of observable reference origins/destinations both found and correctly described; missed navigation remains in the denominator. Motion-negative false-positive rate is the fraction of premarked intervals without navigation incorrectly labeled as navigation. A negative is misclassified when the union of predicted navigation covers at least half its duration; multiple predictions do not count it repeatedly.

Necessary fields and published support. Field-F1 pools counts for actions, target/tool roles, states/-completion, and work areas separately, then averages F1 over categories with reference requirements. A match needs the required meaning, correct role/category, visual support, and tIoU ≥ 0.5. Missing episodes retain all necessary fields as FN. EgoEasy excludes state/completion from this category average, not from truth checking.

For A distinct nonempty core assertions, B contradicted/insufficient assertions, and U unresolved assertions,

$$\mathrm{UFR}_L = \frac{B}{A}, \qquad \mathrm{UFR}_U = \frac{B+U}{A}, \qquad \mathrm{Unresolved} = \frac{U}{A}. \tag{8}$$

If $A = 0$, UFR is N/A: an empty annotation does not receive a zero-error score. Decorative attributes and repeated wording cannot enlarge A.

For Field's lower bound, undecided predictions earn no confirmed matches. For its upper bound, we assume those predictions are supported, rebuild the allowed matches by meaning and role, and match again. True statements not required by the reference remain neutral, and confirmed errors remain errors. The upper score is therefore recomputed, not obtained by adding a fixed amount.

Important operations and observable precision. Let K count distinct reference operations after grouping equivalent timestamps, not the timestamps themselves. A predicted key moment must describe the same operation, instance, and phase. Its time can be within one second of a human timestamp in that group (an *anchor*). Otherwise, it must lie inside the reference episode (one-second edge tolerance), with confirming native frames within one second of the prediction. Falling inside an episode or evidence window alone is insufficient. Each prediction matches at most one group. Event Recall divides confirmed matches by K; its upper bound admits unresolved valid candidates and rematches with the same K. Strict anchor recall permits only matches near human timestamps.

Observable Exact Recall is the fraction of required facts recovered as precisely as the video supports, with one-to-one matching and tIoU ≥ 0.5. Unseen identities and decorative details are excluded. A minimum useful description may already be exact; recovering it does not demonstrate an additional finer distinction. Table 12 separates all exact requirements from those needing a further core distinction.

Usable episodes and remaining review. NEER uses all reference Task/Navigation episodes as its denominator. A qualifying episode has a Seg match, both boundaries within one second, every necessary field/event recovered by that same prediction, and no unsupported/unresolved core assertion or material review. A reference episode requiring human handling cannot qualify. Other predictions cannot supply its missing content. All conditions must hold for the same prediction; NEER is not a measured one-click acceptance rate.

Review span is the union of material-review intervals divided by total video duration. A failed or missing output contributes the complete file duration. Raw review also includes concerns about decorative details. Completion is the fraction of scheduled video–method runs producing valid complete timelines; unknown values or review requests are not failures. Evaluator faults mean scoring is incomplete, not that generation failed.

For independent human diagnosis, a problem is unflagged unless material review contains its verified point or covers at least half its interval. The unflagged fraction divides such problems by all diagnosed important problems; density divides total problems by audited video hours. Zero problems gives density zero and an undefined conditional unflagged fraction, not a fabricated $0/0$ rate.

C.3 What the evaluator sees

Table 15 lists inputs and limits for each scoring stage. The fixed main-model evaluator uses medium thinking, temperature 0, top-p 1, top-k -1, and presence/repetition penalties 0/1. It sees anonymous predictions and reference evidence, not method names, confidence, private reasoning, or generation logs. Evaluation costs are separate.

Table 15: Fixed evaluator access. Image identities bind displayed frames to native timestamps. All ordinary judgment stages use a 16,384-token output limit unless noted.

Stage	Input and recovery rule
Prediction decomposition	One episode per request; 12,288 output tokens; one same-limit mechanical repair.
Joint deduplication	Complete same-episode assertion set; one same-input recovery to 32,768 tokens only for invalid truncation.
Text truth / field matching	At most eight assertions per batch, with shared reference context.
Disputed assertions	At most four per batch; 32 uniform whole-video frames plus 24 native frames per associated episode; timestamp deduplication and a 128-image cap.
Event matching	At most eight key moments per batch; candidates restricted to the reference episode and one-second edges.
Non-anchor verification	At most two pairs; 12 native frames within one second of each prediction and 12 around its nearest anchor. At most 48 images after deduplication.

Matches away from human timestamps need confirming frames near the prediction. Citations must identify displayed frames or exact human reference evidence. If truncation invalidates a judgment, we repeatedly split unanswered items into two batches, preserving their order, complete evidence context, and per-call output limit. Existing valid judgments are not resampled; items still failing individually remain unscored. General evaluation fixes apply uniformly to affected methods, retaining earlier versions and unchanged generation outputs.

C.4 Aggregation, intervals, and primary tests

Let Q_H and Q_E be the scores after pooling HD-EPIC and EgoEasy counts separately, including the applicable within-benchmark field macro-average. The reported aggregate is

$$\bar{Q} = \tfrac{1}{2}(Q_H + Q_E). \tag{9}$$

It does not weight 428 EgoEasy files against 60 HD-EPIC files. Human/resource rates are likewise normalized within each benchmark before equal weighting. If either benchmark has no applicable denominator, the aggregate is N/A rather than silently reweighting the other. Summed raw counts need not reproduce a benchmark-balanced ratio.

Sampling intervals resample whole wearers in HD-EPIC and whole collection-directory groups in EgoEasy, retaining all parent recordings, files, and paired methods. We use 10,000 replicates with seed 20260917, recompute pooled scores, and take the 2.5th/97.5th percentiles of differences. There are 7+11 main-test groups, not 488 independent people. These intervals assume approximate group independence; unverified cross-group identity/site dependence could make them too narrow.

Undecided judgments can change scores even when the sampled videos stay fixed. Let $Q_{a,L}, Q_{a,U}$ be EEA's lower and upper scores, and $Q_{b,L}, Q_{b,U}$ the control's. The possible score difference lies in

$$[Q_{a,L} - Q_{b,U},\ Q_{a,U} - Q_{b,L}]. \tag{10}$$

For UFR, negative differences favor EEA. To include sampling uncertainty as well, we recalculate both endpoints in every shared bootstrap resample. The lower endpoint's 2.5th percentile and upper endpoint's 97.5th percentile form the outer interval. It contains the primary difference interval. Undefined resamples are recorded jointly, not redrawn separately. Removing one group at a time instead measures how much that group influences the result; its range is not a confidence interval.

The primary family contains EEA versus five baselines on Seg, Field lower bound, and UFR upper bound: 15 tests. Two-sided group-level exchange tests enumerate all $2^{18} = 262{,}144$ whole-group method-label exchanges. Holm correction keeps family size 15 and $\alpha = 0.05$. The test assumes approximately independent groups and that method labels within a group can be exchanged under the null hypothesis. Other endpoints provide supporting effects and intervals rather than an expanded search for small p-values. Failed generation remains in the reference denominators, review duration, completion, and consumed cost; no assertions are invented to penalize UFR.

C.5 Human checks of the scoring procedure

Even cross-checked human references do not ensure that the evaluator reads and matches predictions correctly. For each calibration item, one qualified person from each of the three groups first judges

independently with full native-video access, blind to method and model judgment. Cross-checking then produces a final judgment or retained uncertainty; initial votes remain recorded. A person exposed to a video's calibration/reference cannot operate its timed human-study task.

We uniformly sample without replacement 20 assertions per benchmark–method stratum for the main experiment (240 total) and ten for the second backbone (60 total), using seed 20260917 and stable item identities. Strata are weighted equally and sampling ignores correctness or EEA advantage. Every stratum reaches its target, so pooled fractions equal weighted rates.

Table 16 reports four-way agreement and error rates; Table 17 gives their model–human confusion counts. False acceptance is the fraction of model-supported claims humans contradict or find insufficiently supported; false rejection is the fraction of human-supported claims the model definitely rejects. Human-undecided cases are separate, not errors.

Table 16: Representative human calibration. Percentages follow the displayed denominators; false acceptance and rejection have different conditioning sets.

Sample	Agreement	False acceptance	False rejection	Human unresolved
Main	90.83 (218/240)	3.27 (5/153)	3.29 (5/152)	4.58 (11/240)
Second	91.67 (55/60)	5.13 (2/39)	2.63 (1/38)	6.67 (4/60)

Table 17: Confusion counts: rows are model judgments, columns final human judgments. S: supported; C: contradicted; I: insufficient evidence; U: indeterminate.

Model judgment	S	C	I	U
Main backbone				
S	146	2	3	2
C	2	29	2	1
I	3	2	37	2
U	1	1	1	6
Second backbone				
S	37	1	1	0
C	0	7	0	1
I	1	0	9	1
U	0	0	0	2

Neither sample is free of false acceptance or rejection.

Before discussion, 197/240 items have unanimous human labels, 22 have a 2–1 split, and 21 have three different labels. The 720 S/C/I/U votes total 436/103/131/50. The unanimous rate is therefore 82.08%, and observed pair agreement is 0.8513889. Chance agreement from the label proportions is 0.4250887. Correcting for that chance level gives Fleiss' $\kappa = 0.742$, which measures independent annotator agreement, not model accuracy.

Table 18 separately checks difficult assertions, reference extraction, non-anchor evidence, episode-level statement coverage, and broad outputs. These checks do not enlarge the representative-sample denominator. Reference checks sample ten necessary fields and ten event groups per benchmark; event checks sample 20 non-anchor matches per benchmark; decomposition checks sample two episodes per benchmark–method. Broad checks use the fixed sampling rule in Appendix E.2.

Table 18: Additional scoring checks. Error means the block-specific substantive error/invalidity criterion, not four-class disagreement.

Diagnostic block	Items	Errors	Unresolved
Difficult assertions	60	12	7
Reference fields / event groups	40	1	0
Non-anchor event evidence	40	2	1
Episode decomposition coverage	24	1	0
Broad statements	60	3	2

The reference compilation check passes 39/40 items (97.50%); non-anchor evidence is confirmed in 37/40 (92.50%); 23/24 episodes pass decomposition coverage (95.83%). The last rate counts passing episodes, not individual claims recovered. These samples reveal scorer errors; they cannot certify every judgment or exclude shared Qwen generator–evaluator errors.

D COMPLETE WHOLE-VIDEO RESULTS

We unpack the main table's average into benchmark results, then examine undecided judgments, navigation, and computation. Unless stated otherwise, quality/review values are percentages; L/U brackets reflect undecided judgments, not sampling intervals.

D.1 RESULTS WITHIN EACH BENCHMARK

Table 19 reports timing and navigation within each benchmark. Table 20 checks content in those outputs; Table 21 adds exact-detail recovery, usable episodes, review, and completion.

Table 19: Temporal structure by benchmark. Round: both-leg round-trip recall; End: observable endpoint recovery; Motion FP: false-positive rate on premarked non-crossing intervals.

Method	Seg	Boundary	Nav	Round	End	Motion FP
HD-EPIC						
Direct	22.99	12.57	12.74	6.58	12.15	30.10
Dense-Group	33.30	18.14	22.30	14.47	20.33	21.94
Dense-Group+SR	31.40	16.75	21.19	13.16	19.86	20.92
DeMiAn-Ann	20.18	10.27	6.15	2.63	8.88	30.61
VideoHV-Ann	29.41	15.08	21.44	13.16	23.13	21.94
EEA	46.24	39.87	41.52	38.16	44.39	8.67
EgoEasy-Bench						
Direct	16.12	8.50	8.88	3.33	7.14	34.98
Dense-Group	30.30	14.53	18.18	11.25	17.00	25.10
Dense-Group+SR	28.28	14.42	17.77	10.42	16.89	24.07
DeMiAn-Ann	15.10	7.53	3.99	1.25	5.92	32.10
VideoHV-Ann	25.85	12.74	18.33	10.42	19.10	25.10
EEA	42.02	36.05	33.93	30.00	37.49	11.93

Table 20: Core semantics by benchmark, with full semantic [L,U] bounds.

Method	Field [L,U]	UFR [L,U]	Event [L,U]
HD-EPIC			
Direct	[35.77, 38.38]	[22.37, 27.21]	[37.05, 41.97]
Dense-Group	[48.25, 51.09]	[15.72, 19.63]	[53.39, 57.18]
Dense-Group+SR	[50.11, 52.74]	[13.91, 17.84]	[53.79, 57.69]
DeMiAn-Ann	[39.91, 43.84]	[24.86, 30.93]	[41.23, 46.04]
VideoHV-Ann	[49.99, 53.78]	[16.72, 22.68]	[51.58, 56.00]
EEA	[57.51, 60.48]	[7.49, 11.90]	[65.44, 69.85]
EgoEasy-Bench			
Direct	[29.74, 32.78]	[27.18, 33.12]	[30.05, 34.92]
Dense-Group	[44.02, 47.09]	[18.86, 23.74]	[47.18, 51.76]
Dense-Group+SR	[45.04, 47.93]	[16.84, 21.76]	[47.63, 52.16]
DeMiAn-Ann	[33.71, 37.77]	[28.11, 34.05]	[35.11, 39.97]
VideoHV-Ann	[45.93, 49.84]	[19.64, 25.78]	[46.61, 50.95]
EEA	[55.76, 59.12]	[8.62, 12.38]	[63.03, 67.26]

Table 21: Precision retention, usable episodes, and review. Complete is legal-output completion; All review includes decorative concerns; Failed duration is the fraction of video time in failed runs.

Method	Exact	NEER	Review	Complete	All review	Failed duration
HD-EPIC						
Direct	31.12	1.04	25.03	96.67	27.21	3.72
Dense-Group	44.31	3.60	13.15	98.33	15.34	1.36
Dense-Group+SR	46.27	4.18	15.23	96.67	17.42	2.42
DeMiAn-Ann	35.17	1.16	14.06	96.67	16.27	3.06
VideoHV-Ann	49.14	3.36	20.98	96.67	23.40	2.79
EEA	54.55	18.33	23.44	98.33	27.85	1.31
EgoEasy-Bench						
Direct	24.16	0.51	28.02	96.03	30.24	4.07
Dense-Group	38.29	2.13	15.10	98.60	17.36	1.44
Dense-Group+SR	40.18	2.56	17.01	97.90	19.22	2.12
DeMiAn-Ann	28.22	0.59	15.98	96.73	18.19	3.31
VideoHV-Ann	44.20	2.36	22.04	97.20	24.31	2.86
EEA	50.73	8.82	20.84	98.36	25.78	1.62

EEA improves all primary metrics, exact-detail recovery, and NEER in both benchmarks, but does not minimize review. The gains include supported detail, not uniformly vaguer descriptions.

Table 22 breaks Field-F1 down by category, with lower-bound TP/FP/FN and independently re-matched upper bounds. HD-EPIC uses all four categories; EgoEasy uses action, role, and place. EgoEasy's erroneous state assertions remain evaluated by UFR and NEER despite not entering Field-F1.

Table 22: Field decomposition. F1 columns are lower/upper semantic bounds; the first three counts correspond to lower-bound matching.

Category	TP	FP	FN	F1 L	F1 U / state errors
HD-EPIC — Direct					
Action	247	91	613	41.24	43.85
Target/tool role	249	92	783	36.27	38.88
State/completion	122	45	490	31.32	33.93
Place	152	56	528	34.23	36.84
HD-EPIC — Dense-Group					
Action	351	82	509	54.29	57.14
Target/tool role	355	83	677	48.30	51.14
State/completion	186	44	426	44.18	47.02
Place	220	52	460	46.22	49.06
HD-EPIC — Dense-Group+SR					
Action	365	75	495	56.15	58.78
Target/tool role	371	76	661	50.17	52.80
State/completion	195	40	417	46.04	48.67
Place	230	47	450	48.07	50.70
HD-EPIC — DeMiAn-Ann					
Action	285	122	575	44.99	48.91
Target/tool role	298	128	734	40.88	44.80
State/completion	148	63	464	35.97	39.90
Place	176	75	504	37.81	41.73
HD-EPIC — VideoHV-Ann					
Action	375	106	485	55.93	59.71
Target/tool role	389	110	643	50.82	54.60
State/completion	201	57	411	46.21	50.00
Place	229	65	451	47.02	50.80
HD-EPIC — EEA					
Action	429	53	431	63.93	66.92
Target/tool role	449	55	583	58.46	61.43
State/completion	242	30	370	54.75	57.73
Place	256	32	424	52.89	55.86

Table 22 (continued)

Category	TP	FP	FN	F1 L	F1 U / state errors
EgoEasy-Bench — Direct					
Action	589	266	1950	34.71	37.75
Target/tool role	472	213	2128	28.74	31.78
State/completion	–	–	–	N/A	281 erroneous claims
Place	254	115	1349	25.76	28.80
EgoEasy-Bench — Dense-Group					
Action	931	247	1608	50.09	53.16
Target/tool role	769	204	1831	43.05	46.12
State/completion	–	–	–	N/A	198 erroneous claims
Place	414	110	1189	38.93	42.00
EgoEasy-Bench — Dense-Group+SR					
Action	976	229	1563	52.14	55.02
Target/tool role	786	184	1814	44.03	46.91
State/completion	–	–	–	N/A	171 erroneous claims
Place	411	96	1192	38.96	41.84
EgoEasy-Bench — DeMiAn-Ann					
Action	714	352	1825	39.61	43.67
Target/tool role	586	289	2014	33.73	37.79
State/completion	–	–	–	N/A	304 erroneous claims
Place	281	139	1322	27.78	31.84
EgoEasy-Bench — VideoHV-Ann					
Action	1052	332	1487	53.63	57.54
Target/tool role	831	262	1769	45.00	48.91
State/completion	–	–	–	N/A	181 erroneous claims
Place	423	134	1180	39.17	43.08
EgoEasy-Bench — EEA					
Action	1091	149	1448	57.74	61.10
Target/tool role	1059	144	1541	55.69	59.05
State/completion	–	–	–	N/A	74 erroneous claims
Place	622	85	981	53.85	57.21

EEA leads every scored category: the Field improvement is not confined to one type of required fact.

D.2 Do semantic uncertainty or group influence change the conclusions?

Table 23 reports every primary comparison, not only the strongest baseline on each metric. Differences are EEA minus the comparator in percentage points: positive favors EEA for Seg/Field, negative for UFR. Effects use unrounded scores; subtracting two printed cells can differ by 0.01 pp. Table 24 bounds score differences under the allowed judgments on undecided claims. Table 25 adds group-sampling uncertainty and tests Seg after removing each group in turn.

Table 23: All 15 primary comparisons. Effects and 95% paired group-bootstrap intervals are in pp. Holm correction uses the fixed family of 15 tests.

Comparator	Endpoint	Effect	95% CI	Raw p	Holm p
Direct	Seg	$+24.58$	$[+17.76, +31.42]$	0.00006866	0.00075531
Direct	Field_L	$+23.89$	$[+17.48, +31.97]$	0.00004578	0.00059509
Direct	UFR_U	-18.03	$[-25.06, -10.99]$	0.00006104	0.00073242
Dense-Group	Seg	$+12.34$	$[+6.08, +18.57]$	0.00247955	0.01735687
Dense-Group	Field_L	$+10.50$	$[+4.96, +17.82]$	0.00534058	0.02136230
Dense-Group	UFR_U	-9.55	$[-15.18, -4.07]$	0.00343323	0.01739502
Dense-Group+SR	Seg	$+14.29$	$[+7.82, +20.74]$	0.00190735	0.01525879
Dense-Group+SR	Field_L	$+9.06$	$[+3.36, +16.51]$	0.01373291	0.02746582
Dense-Group+SR	UFR_U	-7.67	$[-13.27, -2.04]$	0.00839233	0.02517700
DeMiAn-Ann	Seg	$+26.49$	$[+19.63, +33.36]$	0.00000763	0.00011444
DeMiAn-Ann	Field_L	$+19.83$	$[+13.17, +27.68]$	0.00010681	0.00106812
DeMiAn-Ann	UFR_U	-20.35	$[-27.48, -13.22]$	0.00003052	0.00042725
VideoHV-Ann	Seg	$+16.51$	$[+9.61, +23.40]$	0.00129700	0.01167297
VideoHV-Ann	Field_L	$+8.67$	$[+1.47, +14.85]$	0.01792908	0.02746582
VideoHV-Ann	UFR_U	-12.10	$[-18.04, -6.18]$	0.00289917	0.01739502

Table 24: EEA-minus-baseline ranges under semantic uncertainty (pp), not confidence intervals.

Comparator	Field range	UFR range	Event range
Direct	$[+21.06, +27.05]$	$[-22.11, -12.64]$	$[+25.79, +35.01]$
Dense-Group	$[+7.54, +13.67]$	$[-13.63, -5.16]$	$[+9.76, +18.27]$
Dense-Group+SR	$[+6.31, +12.23]$	$[-11.75, -3.24]$	$[+9.31, +17.85]$
DeMiAn-Ann	$[+15.83, +22.99]$	$[-24.43, -14.35]$	$[+21.23, +30.39]$
VideoHV-Ann	$[+4.83, +11.84]$	$[-16.18, -6.04]$	$[+10.76, +19.46]$

Table 25: Combined semantic/sampling outer intervals and leave-one-group-out Seg influence ranges (pp). Influence ranges are not confidence intervals.

Comparator	Field outer CI	UFR outer CI	Seg group-exclusion range
Direct	$[+15.02, +34.47]$	$[-25.08, -10.17]$	$[+21.46, +27.91]$
Dense-Group	$[+2.61, +20.14]$	$[-15.36, -3.79]$	$[+8.86, +15.62]$
Dense-Group+SR	$[+1.18, +18.76]$	$[-13.44, -1.96]$	$[+10.77, +18.08]$
DeMiAn-Ann	$[+9.48, +31.63]$	$[-27.86, -12.94]$	$[+23.37, +29.84]$
VideoHV-Ann	$[+0.21, +16.08]$	$[-18.31, -5.18]$	$[+12.86, +20.39]$

All 15 primary comparisons favor EEA after Holm correction. Under adverse semantic judgments, Field and Event remain higher and UFR lower. Field/UFR gains also survive combined sampling uncertainty; Seg gains survive each group exclusion. Field/UFR have zero undefined resamples. These tests do not exclude systematic scorer bias.

Table 26 makes unresolved judgments visible. Its assertion counts sum the benchmarks, while the rate averages their separate U/A values equally; therefore the displayed rate need not equal total U divided by total A.

Table 26: Public assertion counts and benchmark-balanced unresolved rates (%).

Method	Assertions A	Unsupported B	Unresolved U	Unresolved rate
Direct	16037	4107	895	5.39
Dense-Group	18047	3219	823	4.39
Dense-Group+SR	17831	2833	820	4.42
DeMiAn-Ann	18529	5013	1107	6.00
VideoHV-Ann	18109	3384	1102	6.05
EEA	18643	1538	740	4.08

EEA publishes the most core claims but the fewest unsupported or undecided claims, ruling out simply saying less.

D.3 Navigation, boundaries, and strict anchor localization

Table 19 separates finding navigation from correctly naming its starting and ending work areas. Across benchmarks, EEA reaches 37.73% navigation F1, 34.08% both-leg round-trip recall, and 40.94% observable endpoint recovery; the respective strongest baselines achieve 20.24%, 12.86%, and 21.12%. Missed navigation still counts against endpoint recovery, so correctly naming only a few easy work areas is insufficient. Boundary-F1 improves from the best baseline's 16.34% to 37.96%.

Motion-negative evaluation uses the 196/486 premarked HD/EgoEasy intervals in Table 12. EEA's balanced false-positive rate is 10.30%, versus the best baseline's 22.50%. This diagnoses the specified camera/object/local-motion negatives, not every background frame in the video.

We also require every event match to be within one second of a human timestamp, retaining the reference-operation denominator. Table 27 compares this strict anchor recall and Task-only Seg.

Table 27: Additional temporal diagnostics on the same saved outputs, with equal benchmark weights (%).

Method	Task-only Seg	Strict anchor recall
Direct	23.45	25.36
Dense-Group	37.83	41.73
Dense-Group+SR	35.25	42.15
DeMiAn-Ann	25.50	31.88
VideoHV-Ann	31.61	40.97
EEA	47.47	47.63

EEA leads both diagnostics: task segmentation improves, and event gains do not depend on crediting times away from human anchors.

D.4 Computation, enhanced Dense-Group, and failures

Table 28 reports computation and failures, including initial annotation, captioning, generated questions, reasoning, observation, final descriptions, repair, and failed attempts. A shared initial annotation is charged to every method that uses it, even when cached; the separate server-use record counts the shared call once. Evaluation remains a separate cost.

Calls, processed input/output tokens, and unique-frame counts are normalized by video hours within each benchmark, then averaged equally. Tokens include visual input but are not exact FLOPs. A unique frame is identified by video identity and native timestamp. Wall/video ratio divides elapsed time by video duration. The weighted p95 is when 95% of files finish under equal benchmark weights, not a per-hour rate. Concurrent call times are not added as GPU hours; monetary cost requires an actual rate.

Table 28: Complete-test resources. Calls, tokens, and frames are per video hour; p95 is seconds; failure counts sum all 488 files.

Method	Calls/h	Tokens/h	Frames/h	Wall/video	p95	Failures
Direct	48.97	896921	1800	0.53	2174	19
Dense-Group	53.71	1874741	3600	1.04	3968	7
Dense-Group+SR	161.14	5782811	3600	2.83	8126	11
DeMiAn-Ann	936.81	4469472	2550	4.17	13142	16
VideoHV-Ann	2454.98	13387536	3900	8.23	30218	14
EEA	592.37	8976432	8427	5.11	19864	8

EEA uses more tokens than Dense-Group but fewer than VideoHV-Ann. Table 29 tests a denser Dense-Group: 60-second/2-fps instead of 120-second/1-fps windows, with unchanged 25% overlap and output limits. All three configurations use the same 60 files; unchanged Dense-Group and Full reuse their original outputs.

A cost reference $B(L) = 382{,}417 + 9{,}284{,}613L$ tokens, with L in video hours, is fitted on eight non-test files per benchmark using the fixed EEA policy. Nonnegative least squares uses equal benchmark weights; the intercept is then raised to cover all 16 measurements. This measures cost without selecting EEA behavior by quality. Tokens/B is the ratio of summed actual tokens to summed $B(L)$ within each benchmark, then equally averaged.

Table 29: One predefined enhanced-resource point. Quality values are percentages; brackets are semantic bounds. The enhanced baseline remains less expensive than EEA.

Configuration	Seg	Field [L,U]	UFR [L,U]	Event [L,U]	Tokens/B	Failures
Dense-Group	32.14	$[46.52, 49.41]$	$[17.38, 21.79]$	$[50.41, 54.63]$	0.214	1
Dense-Group enhanced	36.82	$[50.67, 53.54]$	$[14.91, 18.83]$	$[56.28, 60.42]$	0.587	1
EEA	44.27	$[57.12, 60.28]$	$[7.92, 12.08]$	$[65.08, 69.41]$	0.967	1

Enhanced Dense-Group improves quality but remains cheaper and below EEA at the measured settings. This is a quality–resource comparison, not evidence of superiority at identical tokens or FLOPs. The Uniform component control provides the narrower equal-visual-input test in Appendix E.2.

EEA has eight generation failures among 488 files. Its 98.35% completion rate averages the two benchmark rates, rather than reporting $480/488$. Failures retain reference obligations, full-file review duration, and consumed resources. Legitimate review/unknown outputs are not failures, and evaluator faults are not charged as generation errors. The corresponding Dense-Group completion rate is 98.47%; neither completion rate by itself establishes semantic quality.

E CONTROLLED TESTS OF THE MECHANISMS

We first switch verification and coverage on or off, letting each variant run its own investigations. We then fix Full's investigation sequence and change only the views, labels shown, or descriptions allowed. These tests separate whole-process changes from component changes.

E.1 INDEPENDENT VERIFICATION AND COVERAGE RUNS

Let V indicate whether targeted verification and release checks are enabled, and C whether independent coverage is enabled; 1 means on and 0 off. The four conditions share initial observations, pixels, annotation, and its charged cost, then run independently. Disabled-module budgets are not reassigned, and no variant receives evidence found later by Full.

$V0C0$ reports the initial annotation without applying Full's semantic verification requirements. $V1C0$ checks existing claims but disables both coverage and any equivalent final-stage search for missed events. $V0C1$ uses its own coverage findings, not Full's later verification or accepted facts. All conditions retain source, time, and response-format checks. Full reuses the main-experiment output for that video and seed. Table 30 reports segmentation and semantic bounds; Table 31 isolates each path's effect with the other on or off.

Table 30: Seg and semantic bounds for the seven mechanism conditions. The last three rows use fixed-schedule component replay; NEER, review, and relative tokens appear in Table 2.

Condition	Seg	UFR [L,U]	Event [L,U]
Neither path	31.08	[20.92, 27.36]	[49.73, 55.28]
Verification only	33.17	[10.74, 16.28]	[51.24, 56.83]
Coverage only	38.06	[19.63, 26.11]	[60.87, 66.18]
Full	44.27	[7.92, 12.08]	[65.08, 69.41]
Uniform views	40.11	[11.73, 16.19]	[60.13, 65.22]
Labels visible	42.03	[12.66, 17.31]	[62.51, 67.54]
No broad outputs	44.27	[7.08, 11.14]	[59.14, 64.02]

Table 31: Conditional closure effects (pp), with 95% group-bootstrap intervals below each estimate. UFR uses its upper bound and Event its lower bound.

Comparison	UFR effect	Event effect	NEER effect
	-11.08	$+1.51$	$+4.43$
$V1C0 - V0C0$	$[-15.17, -6.99]$	$[-3.57, +6.59]$	$[+2.17, +6.69]$
	-14.03	$+4.21$	$+6.99$
Full$-V0C1$	$[-18.34, -9.72]$	$[-0.61, +9.03]$	$[+4.57, +9.41]$
	-1.25	$+11.14$	$+3.04$
$V0C1 - V0C0$	$[-5.36, +2.86]$	$[+5.93, +16.35]$	$[+0.74, +5.34]$
	-4.20	$+13.84$	$+5.60$
Full$-V1C0$	$[-8.24, -0.16]$	$[+8.58, +19.10]$	$[+3.17, +8.03]$

Verification reduces unsupported claims whether coverage is on or off, but alone shows no clear Event Recall gain. Coverage improves event recovery whether verification is on or off. Together they produce more episodes that are both correct and complete, as measured by NEER. This shows complementary contributions; it does not establish a greater-than-additive interaction.

E.2 FIXED-SCHEDULE COMPONENT CONTROLS

Each component test follows Full's initial annotation and the same investigation locations, order, questions, and access permissions. It makes decisions from its own observations, not Full's accepted

answers. Only calls with identical inputs can be reused; later views cannot be shown early. Repeating Full unchanged must reproduce its facts and counts. A variant cannot add investigations in response to changed answers, so the results apply to this fixed sequence, not every sequence the variant might choose.

Uniform views: keep new input quantity, change view selection. Each replacement uses the same allowed time interval and image region, preserving image count, resolution, crop size, questions, calls, and output limits. We divide the time interval into equal bins and select the native frame nearest each bin center. Duplicate views are replaced by the nearest unused permitted view, with deterministic timestamp/coordinate tie-breaking. Valid replacement exists for 783/840 investigations (93.21%). If replacement is impossible, that step is omitted from both sides of the local comparison; the complete-video variant retains Full's original step. Unrelated or repeated frames do not fill missing replacements; selection never depends on the outcome.

Labels visible: keep evidence, reveal the proposed answer. The observation call receives the same new views, timestamps, and question, but also sees the proposed labels. No human answers or future evidence are added. This tests whether showing a suggested answer changes the observation.

No broad outputs: preserve exact requirements, withhold broad-only facts. Valid exact descriptions and episode boundaries remain unchanged. Information supported only at a broader level is sent to local review rather than replaced with an exact guess. This rule itself increases review, so that increase alone cannot establish usefulness. We also check how much information is lost and whether humans confirm the withheld descriptions. Table 32 compares Full with each single-component control.

Table 32: Component effects (pp), with 95% intervals. Field/Event use lower bounds, UFR its upper bound. Fixed-schedule controls have the scope described above.

Comparison	Field	UFR	Event	Review
Full − Uniform	$+4.37$ $[+1.18, +7.56]$	-4.11 $[-7.56, -0.66]$	$+4.95$ $[+1.31, +8.59]$	-4.62 $[-8.97, -0.27]$
Full − Labels visible	$+2.84$ $[-0.39, +6.07]$	-5.23 $[-8.77, -1.69]$	$+2.57$ $[-1.02, +6.16]$	-1.85 $[-6.19, +2.49]$
Full − No broad outputs	$+5.18$ $[+1.42, +8.94]$	$+0.94$ $[-2.48, +4.36]$	$+5.94$ $[+2.21, +9.67]$	-15.78 $[-20.31, -11.25]$

Choosing targeted rather than uniform views improves fact and event recovery while reducing unsupported claims. Hiding labels reduces unsupported claims, but does not clearly improve overall Field or Event scores. Keeping broad descriptions preserves more content than NoBroad; their UFR difference remains uncertain. Human time is measured separately.

Check anchoring where the initial candidate is right and wrong. Table 33 compares label-hidden and label-visible outcomes for 520 correct and 250 wrong initial candidates; 70 remain undecided. A final choice counts as correct only if the annotation retains a supported statement containing the required information. Review or unknown alone earns no recovery credit, but is not a false statement. Human-undecided cases remain outside the correct/wrong comparison.

Table 33: Candidate-stratified correctness under identical evidence. Differences are Full minus Labels visible, in pp [95% CI].

Initial candidate	Full	Labels visible	Difference
Correct	486/520 (93.46%)	488/520 (93.85%)	-0.38 $[-2.47, +1.70]$
Incorrect	183/250 (73.20%)	151/250 (60.40%)	$+12.80$ $[+6.94, +18.66]$

Hiding labels helps mainly when the initial suggestion is wrong, consistent with less influence from that suggestion rather than improvement in every judgment. For broad outputs, fixed video/fact hashes select up to 30 removed facts per benchmark. Human checks confirm 55/60 as supported and useful, find three substantive errors, and leave two unresolved. The principal rate is 55/60 (91.67%),

not the resolved-only 55/58 (94.83%). This supports useful information retention, not error-free broad descriptions.

F HUMAN WORK AND FINAL ANNOTATION QUALITY

F.1 ASSIGNMENTS, VIDEO ACCESS, AND COMPLETE TIMING

Each of 36 preselected files enters three workflows: scratch annotation, correction of development-selected Dense-Group, and correction of EEA. The 108 assignments are not independent videos. Correction uses the original outputs, including failures requiring reconstruction.

Thirty-six operators come from the three-group pool. Seeded video ordering and cyclic Latin rotation assign groups to workflows in turn, giving each group six assignments per benchmark–workflow combination. Qualification, experience, and duration are balanced before tasks are opened. No person works on one video in multiple conditions or has previously made/reviewed its gold or participated in its scorer calibration.

The interface, task manual, full-video playback, timeline editing, evidence-link functions, and submission mechanism are shared. Correction hides method identity and internal confidence, while scratch visibly starts empty; complete operator blinding across scratch/correction is therefore not claimed. Production reviewers see anonymous final outputs and video, not workflow labels or hidden gold. Only the separate research scorer can use reference answers.

Active time includes viewing, reading, replay, omission search, boundary/field editing, event addition, and rework. Losing focus pauses the timer; 30 seconds without input prompts confirmation rather than deleting time spent watching. One qualified reviewer from each other group independently checks the complete source/output before cross-checking. Their active review, disagreement resolution, repeat acceptance, and operator rework are all added. Simultaneous workers contribute separate person-time. Breaks, technical waiting, and additional research calibration are excluded; active work repeated after a technical fault remains included.

For video i in benchmark d, let $T_{w,d,i}$ be total active person-minutes under workflow w, and $L_{d,i}$ the video hours. With H and E denoting HD-EPIC and EEB in the subscripts, we compute

$$H_{w,d} = \frac{\sum_i T_{w,d,i} \text{ (active person-minutes)}}{\sum_i L_{d,i} \text{ (video hours)}}, \qquad H_w = \tfrac{1}{2}(H_{w,H} + H_{w,E}). \tag{11}$$

Savings against control b are $100(1 - H_{\mathrm{EEA}}/H_b)$, not an average of individual percentage savings.

A six-file non-test pilot, three files per benchmark selected by duration stratum and fixed hash, fixes $C = 1440$ person-minutes/video-hour: twice the largest observed scratch rate. Every workflow has the same cap $\max(30, CL_i)$ person-minutes, including review/rework, where L_i is video hours. An unfinished assignment establishes only that completion takes longer than the time spent (right-censoring), not an invented completed time. No formal assignment is censored. Table 34 separates operation and review/rework time and reports completion and rework rounds.

Table 34: Active human work under equal benchmark weights. Time is person-minutes/video-hour; rework rounds are recorded totals.

Workflow	Operation	Review/rework	Total H	Accepted	Rework rounds
Scratch	337.8	183.6	521.4	36/36	35
Correct Dense-Group	268.9	162.7	431.6	36/36	24
Correct EEA	151.7	141.2	292.9	36/36	13

EEA reduces both operation and review/rework time, with every assignment completed. Acceptance requires no substantive error or omission found in review; truly unidentifiable facts may stay unknown. Acceptance is not perfect research scoring.

F.2 THE PRESPECIFIED AGGREGATE-QUALITY REQUIREMENT

Before testing, we set limits on benchmark-averaged quality: declines below 2 pp in Seg/Field/Event and a UFR increase below 1 pp. These limits are not selected from EEA's results; the pilot checks workflow and timing, not which limits EEA passes.

For EEA-minus-control differences, the 95% interval must have a lower endpoint above -2 pp for Seg/Field/Event and an upper endpoint below $+1$ pp for UFR. Undecided judgments are least favorable to EEA: $\mathrm{EEA}_L - \mathrm{control}_U$ for Field/Event, $\mathrm{EEA}_U - \mathrm{control}_L$ for UFR. All four conditions must pass against both controls. A nonsignificant difference alone cannot establish non-inferiority. Group resampling retains all three workflows together.

All workflows share the same 36-file references: HD-EPIC contains 271 episodes and 535 event groups; EgoEasy contains 105 episodes and 154 groups. HD-EPIC requires 270 action, 324 role, 193 state, and 214 place facts. EgoEasy requires 105 action, 107 role, and 67 place facts. Table 35 gives final quality by benchmark and on average; Table 36 tests the quality limits; Table 37 reports human-time savings and their intervals.

Table 35: Final human-workflow quality (%). Brackets are semantic [L,U] bounds, not confidence intervals. Benchmark rows are the same outputs as the aggregate rows.

Workflow	Seg	Field [L,U]	UFR [L,U]	Event [L,U]
HD-EPIC				
Scratch	95.97	[97.14, 97.14]	[0.33, 0.64]	[91.78, 92.52]
Correct Dense-Group	96.41	[95.77, 95.77]	[0.86, 1.33]	[90.28, 90.47]
Correct EEA	93.90	[92.93, 93.27]	[0.39, 0.75]	[99.25, 99.44]
EgoEasy-Bench				
Scratch	89.69	[93.13, 94.02]	[2.43, 3.00]	[98.05, 98.05]
Correct Dense-Group	88.31	[93.77, 94.65]	[2.36, 2.79]	[98.70, 99.35]
Correct EEA	91.48	[97.13, 97.69]	[2.57, 3.07]	[90.26, 90.91]
Equal benchmark weights				
Scratch	92.83	[95.14, 95.58]	[1.38, 1.82]	[94.91, 95.29]
Correct Dense-Group	92.36	[94.77, 95.21]	[1.61, 2.06]	[94.49, 94.91]
Correct EEA	92.69	[95.03, 95.48]	[1.48, 1.91]	[94.76, 95.17]

Table 36: EEA-minus-control quality differences, with 95% intervals, in pp. Semantic assignments are least favorable to EEA; all eight aggregate criteria pass.

Control	Seg	Field	Event	UFR
Scratch	-0.14 $[-1.18, +0.90]$	-0.55 $[-1.46, +0.36]$	-0.53 $[-1.48, +0.42]$	$+0.53$ $[+0.13, +0.92]$
Correct Dense-Group	$+0.33$ $[-0.72, +1.38]$	-0.18 $[-1.10, +0.74]$	-0.15 $[-1.12, +0.82]$	$+0.30$ $[-0.10, +0.70]$

Table 37: Human savings under the prespecified aggregate-quality requirement, with 95% intervals. Every workflow completes all assignments without censoring.

Control	Person-minutes/video-hour difference	Saving (%)
Scratch	-228.5 $[-272.8, -184.1]$	43.82 [36.9, 50.7]
Correct Dense-Group	-138.7 $[-173.2, -104.6]$	32.14 [23.8, 40.4]

All eight average-quality requirements pass, and both human-time saving intervals exclude zero. EEA reduces measured work within these preset quality limits, not with better-than-expert quality or success on each benchmark separately.

Why the benchmark breakdown matters. The limits do not hold within each benchmark. Under judgments least favorable to EEA, differences against scratch are -2.07 pp in HD-EPIC Seg, -4.21 pp in HD-EPIC Field, and -7.79 pp in EEB Event. Against corrected Dense-Group, they are -2.51, -2.84, and -9.09 pp. Gains elsewhere offset these excessive losses in the predefined average; this is not a pass within each benchmark.

The intervals concern this personnel allocation and grouping. Operators appearing across different videos may introduce additional dependence; these data do not by themselves estimate performance for an arbitrary annotator population. Model computation, research evaluation, and active human work remain separately accounted for, so human savings are not automatically total economic savings.

F.3 Do review flags locate the problems that remain?

Independent blind diagnosis checks initial EEA/Dense-Group outputs on the same 36 files without showing findings to timed operators. It deduplicates substantive fact errors, omissions, and boundary problems using Appendix C.2. Table 38 reports their number and the fraction outside marked review.

Table 38: Important problems in the initial human-study outputs. Density uses the actual 7.319 video hours, not equal benchmark weighting.

Initial output	Problems	Unflagged	Unflagged (%)	Problems/video-hour
Dense-Group	218	171	78.44	29.79
EEA	81	20	24.69	11.07

EEA leaves fewer substantive problems, and fewer of them lack a review mark. People therefore face fewer problems to fix or find without guidance, consistent with the measured savings despite more marked review time. This separate diagnosis does not prove which change caused the savings; it covers only the 36 human-study files.

G Additional backbone and sensitivity analyses

We next ask whether the main improvements remain with another fixed model, whether they change across three seeds, and whether scoring thresholds or video length explain them. These checks use the existing videos and predefined subsets, not new benchmarks selected after viewing results.

G.1 A second fixed Qwen backbone

All generative roles switch to Qwen3-VL-30B-A3B-Instruct on the same 60 files, without selecting prompts or settings from test results. EEA runs independently without reusing main-backbone outputs. Video access and references remain identical. The evaluator stays fixed so changing the generator does not change scoring; Appendix C.5 reports separate human calibration.

The HD-EPIC subset contains 451 episodes and 926 core events; EgoEasy contains 175 episodes and 262 events. HD-EPIC requires 450 action, 540 role, 321 state, and 356 place facts. EgoEasy requires 175 action, 179 role, and 111 place facts. The mechanism and enhanced-resource comparisons use these same denominators, retaining obligations from failed runs. Table 39 reports second-backbone quality; Table 40 gives paired gains and intervals; Table 41 records tokens, time, and failures.

Table 39: Second-backbone results (%). Semantic bounds are [L,U]. All methods replace every generative role, including perception and readout.

Method	Seg	Field [L,U]	UFR [L,U]	Event [L,U]	NEER	Complete
HD-EPIC						
DeMiAn-Ann	18.42	[39.74, 43.81]	[27.36, 33.12]	[40.82, 46.22]	1.11	96.67
VideoHV-Ann	28.14	[50.82, 54.93]	[18.26, 24.08]	[50.76, 56.16]	3.33	96.67
EEA	40.76	[56.41, 61.19]	[10.63, 15.08]	[61.77, 67.39]	14.19	96.67
EgoEasy-Bench						
DeMiAn-Ann	14.73	[34.39, 38.24]	[29.81, 35.72]	[35.50, 40.46]	0.57	96.67
VideoHV-Ann	25.18	[46.47, 50.54]	[20.91, 26.77]	[44.27, 50.00]	1.71	96.67
EEA	35.64	[52.13, 57.02]	[13.19, 18.71]	[57.25, 62.60]	7.43	100.00
Equal benchmark weights						
DeMiAn-Ann	16.58	[37.06, 41.03]	[28.59, 34.42]	[38.16, 43.34]	0.84	96.67
VideoHV-Ann	26.66	[48.64, 52.73]	[19.59, 25.42]	[47.52, 53.08]	2.52	96.67
EEA	38.20	[54.27, 59.11]	[11.91, 16.89]	[59.51, 64.99]	10.81	98.33

Table 40: Second-backbone EEA-minus-control effects, in pp [95% CI]. Field uses its lower bound and UFR its upper bound.

Comparator	Seg	Field	UFR
DeMiAn-Ann	$+21.62$ $[+14.62, +28.63]$	$+17.21$ $[+9.94, +24.47]$	-17.53 $[-24.01, -11.03]$
VideoHV-Ann	$+11.54$ $[+4.42, +18.66]$	$+5.63$ $[+0.62, +10.65]$	-8.53 $[-14.05, -3.02]$

Table 41: Second-backbone resource use under equal benchmark weighting. Failure counts sum the 60 files.

Method	Tokens/video-hour	Wall/video ratio	Failures
DeMiAn-Ann	5137842	6.4	2
VideoHV-Ann	12684391	16.7	2
EEA	9148576	10.8	1

All three primary gains hold in both benchmarks, with aggregate intervals excluding zero. EEA uses fewer tokens and less time than VideoHV-Ann, but more than DeMiAn-Ann. This tests another Qwen backbone with a Qwen evaluator, not unrelated families. CPU offload prevents attributing cross-backbone timing to the algorithm.

G.2 Three fixed seeds, with benchmark-level variation retained

EEA and Dense-Group use the same 12 files under seeds 20260916/20260917/20260918; first-seed outputs are reused. These 72 outputs are not 72 independent files. Every seed uses 90 episodes and 185 events in HD-EPIC, and 35 episodes and 51 events in EgoEasy. Table 42 lists the three benchmark-averaged scores, their mean, and sample standard deviation (divisor $3-1$). This deviation is neither a confidence interval nor an unseen-seed guarantee.

Table 42: Repeated seeds. Scores and means are percentages; all standard deviations (SD) are pp. Aggregate SD is computed after averaging benchmarks, not before.

Metric	Three seed scores	Mean	Aggregate SD	HD SD	EgoEasy SD
EEA					
Seg	44.92 / 43.41 / 44.48	44.27	0.78	0.73	2.28
Field_L	57.64 / 56.71 / 57.16	57.17	0.47	1.67	1.59
UFR_U	11.74 / 12.35 / 12.15	12.08	0.31	5.40	6.00
Event_L	65.62 / 64.41 / 65.21	65.08	0.62	8.70	7.84
Dense-Group					
Seg	32.76 / 31.48 / 32.18	32.14	0.64	2.19	0.94
Field_L	46.94 / 46.08 / 46.55	46.52	0.43	0.56	1.25
UFR_U	21.43 / 22.12 / 21.82	21.79	0.35	10.16	9.47
Event_L	51.07 / 49.72 / 50.44	50.41	0.68	38.57	38.49

The average score changes little across seeds, but each benchmark can change much more. Table 43 shows the event counts: increases on one benchmark can offset decreases on the other. A stable average therefore does not establish stability within each benchmark.

Table 43: Event matches and recalls for the same repeated runs. The two benchmark event denominators remain 185 and 51.

Method/seed	HD matches	EgoEasy matches	HD (%)	EgoEasy (%)	Aggregate (%)
EEA/20260916	163/185	22/51	88.11	43.14	65.62
EEA/20260917	144/185	26/51	77.84	50.98	64.41
EEA/20260918	176/185	18/51	95.14	35.29	65.21
Dense/20260916	62/185	35/51	33.51	68.63	51.07
Dense/20260917	57/185	35/51	30.81	68.63	49.72

Table 43 (continued)

Method/seed	HD matches	EgoEasy matches	HD (%)	EgoEasy (%)	Aggregate (%)
Dense/20260918	183/185	1/51	98.92	1.96	50.44

EEA's aggregate results remain better across these three listed seeds. However, the two benchmark scores often move in opposite directions: Dense-Group's third-seed Event Recall is 98.92% on HD-EPIC and 1.96% on EgoEasy. Thus, the three averaged scores are similar, but results within each benchmark are not stable.

G.3 Temporal thresholds and duration strata

Table 44 rescores the same outputs at tIoU 0.3/0.5/0.7 and boundary tolerance 0.5/1/2 seconds. Prediction/reference counts remain fixed; stricter overlap thresholds cannot increase matches, and wider boundary tolerances cannot decrease them.

Table 44: Temporal sensitivity (%). Neither predictions nor references are regenerated.

Method	Seg at tIoU 0.3 / 0.5 / 0.7	Boundary at 0.5 / 1 / 2 s
Direct	30.42 / 19.55 / 10.71	6.89 / 10.54 / 17.44
Dense-Group	42.37 / 31.80 / 20.46	10.53 / 16.34 / 24.98
Dense-Group+SR	40.13 / 29.84 / 18.91	9.96 / 15.59 / 24.07
DeMiAn-Ann	27.46 / 17.64 / 9.18	5.76 / 8.90 / 14.55
VideoHV-Ann	38.51 / 27.63 / 17.44	8.53 / 13.91 / 21.64
EEA	56.28 / 44.13 / 31.76	30.09 / 37.96 / 48.43

EEA leads at these thresholds and tolerances, not only the defaults. Table 45 compares EEA with development-selected Dense-Group across duration strata fixed before prediction. Within each benchmark, videos are sorted by duration and identifier into nearly equal groups, assigning remainders to short then middle. Each group pools its own counts; the comparator is not reselected.

Table 45: Duration-stratum EEA-minus-Dense-Group effects in pp [95% CI]. Duration ranges are seconds; strata are formed separately within each benchmark.

Stratum	Files	Duration range (s)	Seg effect	Event effect
HD-EPIC				
Short	20	322–716	$+10.4$ $[+3.1, +17.7]$	$+9.6$ $[+2.2, +17.0]$
Middle	20	724–1118	$+12.8$ $[+5.0, +20.6]$	$+12.0$ $[+4.1, +19.9]$
Long	20	1124–2638	$+15.7$ $[+7.2, +24.2]$	$+14.5$ $[+5.9, +23.1]$
EgoEasy-Bench				
Short	143	42.814989–300.086733	$+10.9$ $[+4.8, +17.0]$	$+14.6$ $[+8.1, +21.1]$
Middle	143	300.086900–300.156811	$+11.7$ $[+4.9, +18.5]$	$+15.7$ $[+8.4, +23.0]$
Long	142	300.157044–300.389633	$+12.6$ $[+4.7, +20.5]$	$+17.2$ $[+8.6, +25.8]$

Advantages persist in all listed strata. HD-EPIC's substantially different durations support the longer-recording result. EgoEasy's middle and long strata differ only by fractions of a second around five minutes; they do not demonstrate scaling to longer continuous sessions.

G.4 Interpretation and reproducibility scope

Together, the experiments support better task–navigation annotations, distinct verification/coverage contributions, and less human work within the specified average-quality limits. Yet 13.58% NEER is not complete automation, and finite viewing can miss events. Quality changes and seed variation within each benchmark remain relevant.

Reproducing these scores requires the fixed videos and splits, human references, extracted scoring references, original predictions, evaluation evidence, per-video counts, and shared resamples. Human-time results also require assignment, prior-exposure, timing, review, and rework records. Code tests or summary tables cannot replace these records. Reproduction must retain the stated failure, uncertainty, averaging, and component-test rules, not infer counts from rounded scores.