

KnowMeBenchV2: Evidence-Grounded Person-Centric Long-Video Understanding

Tingyu Wu^{1,2,3} Guangyu Cao^{2,3} Qizhen Lan^{3,4} Zhisheng Chen^{2,3} Ziyang Weng⁵
Bingkun Zhu⁶ Miao Su² Chenglong Li⁶ Zhengwei Xie⁷ Huacan Wang^{2,3}
Sen Hu^{3,8} Haocun Ye² Nan An⁹ Zaoqu Liu^{1,3,*} Ronghao Chen^{1,3,8,*}
¹IAPM (Guangdong) ²UCAS ³QuantaAlpha ⁴UTHealth ⁵CITYU-DG
⁶THU ⁷USTC ⁸PKU ⁹USYD

*Corresponding authors: zaoqu.liu@iapm.com, ronghao.chen@iapm.com

Abstract

Questions about a recurring person in long videos are rarely answered from a single moment. They require evidence that is sufficient for the claim, comparisons across episodes, synthesis over recurring audiovisual patterns, and refusal when the source material does not support an answer. Existing long-video benchmarks do not isolate these requirements. We introduce KnowMeBenchV2, an evaluation-only benchmark over five public multi-episode video series, with 327 videos and 3,061 questions covering four evidence scopes: single-clip extraction, temporal and event relations, cross-episode ordering, and cross-episode persona synthesis. A separate grounded-abstention task serves as a contrastive support check: its questions are surface-similar to supported person-linked questions but are not licensed by the source evidence. The most interpretive synthesis items are human-authored and quality-audited by trained annotators with full source access. A matched human reference reaches 78.0 on cross-episode synthesis. To diagnose failures, we compare evidence routes under a fixed Qwen3.5-9B answerer and a task-matched pre-tokenization character routing budget. ASR-Retrieval reaches 2.7 on synthesis, while KnowMe reaches 64.9. In this diagnostic, the final answerer reasons over textualized evidence constructed from audiovisual sources rather than raw video tokens. The gap therefore isolates the importance of audiovisual evidence construction and cross-episode organization before answer generation.

1 Introduction

Lifelong digital companions must build evidence-grounded accounts of a person: what someone tends to do, how they respond in context, which patterns recur, and when the evidence is too weak for a claim. KnowMe-Bench (Wu et al., 2026) framed this as person understanding over autobiographical experience and showed that factual recall is only

one part of personal memory; models must also infer subjective states and higher-level principles from time-anchored evidence. That benchmark kept the substrate textual, however, distilling every person-level signal into written autobiographical narrative before evaluation.

Personal experience is not fully captured by transcripts. A preference may surface only through silent choices made across episodes; a habit may appear as a recurring object, gesture, or setting; an emotional tendency may be visible in a facial reaction long before it is articulated. Video and character-centric work shows that important evidence may lie outside the transcript (Zhou et al., 2024; Meng et al., 2025; Song et al., 2024; Zhang et al., 2025), and evidence-grounded video reasoning treats the visual track as a primary, not auxiliary, source of evidence (Wei et al., 2026; Luo et al., 2025; Yuan et al., 2025). KnowMeBenchV2 is the audiovisual successor to KnowMe-Bench: it inherits the evidence-grounded view of person understanding but lifts the substrate from autobiographical text to multi-episode video so that visual, auditory, and temporal cues remain first-class.

No existing benchmark answers whether systems can form such accounts from audiovisual evidence across episodes. Long-video QA scores answer correctness without isolating person-linked evidence sufficiency or cross-episode persona synthesis (Fu et al., 2025; Zhou et al., 2025; Wu et al., 2024; Li et al., 2024; Mangalam et al., 2023; Chandrasegaran et al., 2024; Wang et al., 2024; Ataallah et al., 2025; Ma et al., 2026); narrative and human-centric benchmarks reason about plots, characters, or emotions without treating audiovisual evidence as accumulating personal memory (Tapaswi et al., 2016; Lei et al., 2018; Zhang et al., 2025; Song et al., 2024; Zhou et al., 2024; Meng et al., 2025); long-term memory benchmarks operate over dialogue or text rather than continuous audiovisual signal (Maharana et al., 2024; Wu et al., 2025; Hu

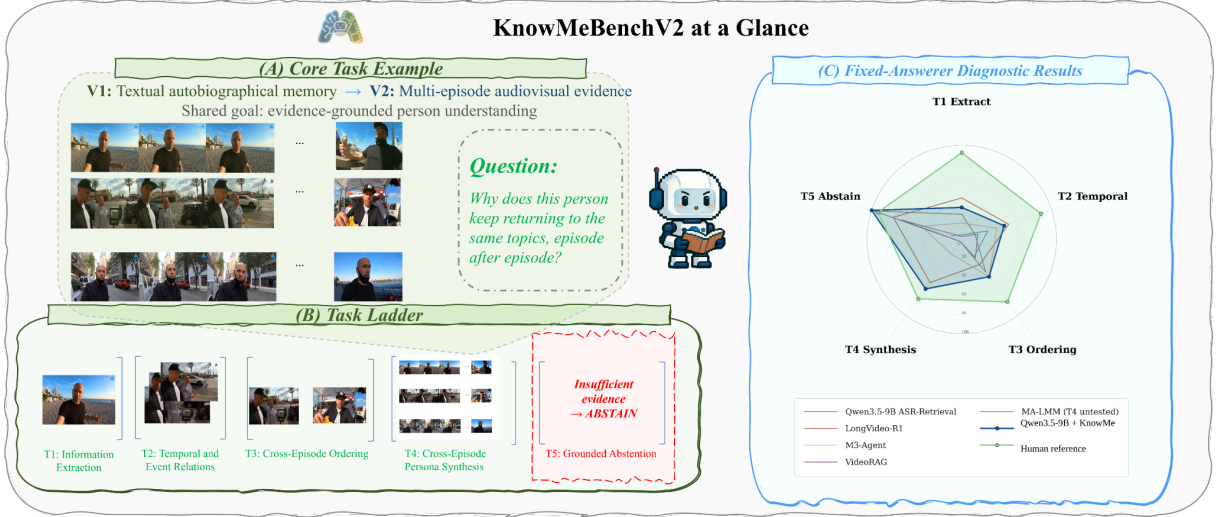


Figure 1: **KnowMeBenchV2 at a glance.** (A) The benchmark extends KnowMe-Bench from textual autobiographical memory to multi-episode audiovisual evidence while preserving the shared goal of evidence-grounded person understanding. (B) T1–T4 form an evidence-scope ladder from localized extraction to cross-episode persona synthesis; T5 separately tests grounded abstention. (C) Fixed-answerer diagnostic results with representative systems and the human reference.

et al., 2026). We therefore ask: **(RQ1)** can current systems form grounded person-level accounts from audiovisual evidence at scopes from single clip to full series; **(RQ2)** how much do failures depend on the evidence representation supplied to the answerer; and **(RQ3)** how large is the gap to a matched human reference?

A benchmark that answers these questions should satisfy four design constraints: **(D1)** evidence scope graded from single clip to full series so the failure point can be localized; **(D2)** person-centric content over recurring subjects rather than generic long video; **(D3)** contrastive grounded abstention, where unsupported items are surface-similar to supported person-linked questions and test evidence-boundary recognition; and **(D4)** a matched human reference collected under the same source-identifier interface, prompts, and rubric as systems. We instantiate these in KnowMeBenchV2, an evaluation-only benchmark for audiovisual personal memory, comprising five public multi-episode video series, 327 source videos, and 3,061 scored questions. The main tasks form a four-step evidence-scope ladder addressing D1–D2: information extraction (T1), temporal and event relations (T2), cross-episode ordering (T3), and cross-episode persona synthesis (T4); a separate grounded-abstention task (T5) addresses D3 and is reported on its own axis.

The most interpretive part of KnowMeBenchV2 is built around human expertise. T4 synthesis

items are human-authored by trained annotators who watch complete source series and write questions whose answers require evidence from multiple separately recorded videos. Human audit checks source support, answerability, cross-episode scope, near-duplicate items, and over-interpretation risk. For the human reference, annotators answer questions on source series they did not author, under the same source-identifier interface and scoring protocol used for systems. This design makes the benchmark auditable while keeping the person-level claims tied to observable audiovisual evidence.

To answer RQ2, we compare evidence routes under a fixed Qwen3.5-9B answerer and a task-matched pre-tokenization character routing budget. The automatic-speech-recognition retrieval route (ASR-Retrieval) supplies retrieved transcript chunks, while KnowMe supplies routed textual summaries constructed from audiovisual evidence. The final answerer does not consume raw video tokens in this diagnostic; it reasons over textualized audiovisual evidence. This setting therefore does not test a new native-video architecture, but measures how performance changes when evidence construction and organization change. On cross-episode persona synthesis, ASR-Retrieval reaches 2.7, while KnowMe reaches 64.9 under the same answerer; the matched human reference is 78.0. The same analysis also exposes an ordering trade-off: ASR-Flat-32K remains stronger on T3, show-

ing that hierarchical compression can lose rank-sensitive details even when it improves synthesis.

In summary, we make four contributions:

- **Benchmark.** KnowMeBenchV2 is an audiovisual successor to KnowMe-Bench (Wu et al., 2026) that evaluates evidence-grounded recurring-person understanding over five public multi-episode video series.
- **Evidence-scope tasks and abstention.** We define a four-step evidence-scope ladder from localized extraction to cross-episode synthesis, plus a grounded-abstention task that serves as a contrastive support check for open-form person-centric evaluation.
- **Human-authored synthesis and quality audit.** We construct the most interpretive synthesis items through trained annotators with full source-series access, source-support checks, over-interpretation filtering, and a matched human reference.
- **Diagnostic findings.** We evaluate representative long-video systems and fixed-answerer evidence routes, showing that local video grounding and cross-episode evidence organization are distinct bottlenecks. Under the same answerer and task-matched pre-tokenization routing budget, structured evidence raises T4 synthesis from 2.7 to 64.9, while T3 exposes an ordering cost from hierarchical compression.

2 Related Work

Long-video and agentic video benchmarks. Long-video QA benchmarks have expanded from ActivityNet-QA and EgoSchema to broader suites covering diverse domains, durations, and task formats (Yu et al., 2019; Mangalam et al., 2023; Fu et al., 2025; Zhou et al., 2025; Li et al., 2024; Wu et al., 2024; Chandrasegaran et al., 2024; Wang et al., 2024; Ataallah et al., 2025). Recent work further moves toward open-ended or agentic evaluation beyond closed-form answering (Ma et al., 2026). These resources stress long-horizon comprehension and realistic evaluation, but do not make cross-episode person-level evidence sufficiency the primary target.

Narrative, character, and human-centric video. MovieQA, TVQA, MovieChat, and SeriesBench

cover story-level or narrative understanding in movies, TV, and drama series (Tapaswi et al., 2016; Lei et al., 2018; Song et al., 2024; Zhang et al., 2025). HumanVBench and VideoForest/CrossVideoQA move closer to person-level evaluation through human-centric perception or person-anchored cross-video QA (Zhou et al., 2024; Meng et al., 2025). KnowMeBenchV2 differs by asking whether systems can synthesize stable behavioral regularities across episodes, including recurring preferences, values, and aversions, under explicit evidence-sufficiency and abstention constraints.

Evidence grounding, temporal localization, and abstention. Temporal localization and moment-retrieval work grounds language queries in video intervals (Gao et al., 2017), while recent long-video reasoning work treats visual evidence as an evaluation or training target (Wei et al., 2026; Luo et al., 2025; Yuan et al., 2025). Separately, abstention and selective prediction improve reliability in language and vision-language QA (Dancette et al., 2023; Srinivasan et al., 2024; Khan and Fu, 2024; Wen et al., 2025), and hallucination evaluation highlights the risk of unsupported visual claims (Li et al., 2023). KnowMeBenchV2 combines these threads: systems must decide whether observable audiovisual evidence is sufficient to support claims about a person across videos, and T5 uses surface-similar unsupported questions as a contrastive support check for open-form evaluation.

Personalized memory and retrieval-based long-video systems. Retrieval-augmented generation, graph/hierarchical retrieval, and agentic memory systems provide mechanisms for organizing retrieved evidence and persistent memory (Lewis et al., 2020; Edge et al., 2024; Zhong et al., 2024; Packer et al., 2023; Park et al., 2023; Hou et al., 2026a; Chen et al., 2026; Hou et al., 2026b). Long-video systems instantiate these ideas through memory augmentation, retrieval, or navigation (He et al., 2024; Ren et al., 2026; Long et al., 2026; Qiu et al., 2026). Long-term and personalized-memory benchmarks evaluate recall, temporal reasoning, memory updating, and person understanding in dialogue histories or personal traces (Maharana et al., 2024; Wu et al., 2025, 2026; Hu et al., 2026). Complementarily, temporal-memory methods explicitly model semantic occurrence time and durative memory for personalized agents (Su et al., 2026). KnowMeBenchV2 differs by making audiovisual

Benchmark / resource	Year	Substrate / scale	P (Pers.)	M (Multi.)	G (Grnd.)	X (Cross.)
<i>Long-video and multimodal video QA</i>						
ActivityNet-QA / EgoSchema	2019–23	Web and egocentric videos	●	●	●	●
Video-MME / MLVU / MVBench	2024–25	Broad video suites	●	●	●	●
LongVideoBench / LVBench / InfiniBench	2024–25	Long public videos, movies, TV	●	●	●	●
VideoEval-Pro	2026	Long videos	●	●	●	●
<i>Narrative, character, and human-centric video</i>						
MovieQA / TVQA	2016–18	Movies and TV episodes	●	●	●	●
MovieChat-1K / SeriesBench	2024–25	Movies and drama series	●	●	●	●
HumanVBench	2024	Human-centric videos	●	●	●	●
VideoForest / CrossVideoQA	2025	Distributed video streams	●	●	●	●
<i>Longitudinal memory and persona resources</i>						
LoCoMo / LongMemEval	2024–25	Dialogue histories	●	●	●	●
KnowMe-Bench	2026	Lifelong digital companion memories	●	●	●	●
CloneMem	2026	Personal digital traces	●	●	●	●
KnowMeBenchV2 (ours)	2026	327 videos / 3,061 Qs	●	●	●	●
<i>Legend: ● explicit ● limited / implicit ● absent</i>						

Table 1: Positioning of KnowMeBenchV2 relative to representative long-video, narrative/character, human-centric, and longitudinal-memory resources. P (Pers.) = person-centric evaluation target; M (Multi.) = multimodal or multi-source evidence; G (Grnd.) = items require explicit evidence pointers, such as span anchors or cited evidence, rather than free-form QA only; X (Cross.) = cross-episode or cross-session synthesis.

evidence construction observable in evaluation: the same benchmark separates local grounding, cross-episode ordering, synthesis, and abstention, and the experiments use fixed-answerer route comparisons to analyze where evidence organization matters. Table 1 summarizes this positioning across person-centric evaluation, multimodal evidence, grounding, and cross-episode synthesis.

3 The KnowMeBenchV2 Benchmark

KnowMeBenchV2 evaluates person-centric long-video understanding over recurring subjects in multi-episode audiovisual sources. Each benchmark instance asks whether a system can construct sufficient evidence from a source set and use it to answer, order, synthesize, or reject a person-linked claim. A source set consists of multiple videos drawn from the same public series, where the same people, settings, activities, and interaction contexts recur across episodes. This setting changes the evidence substrate of person understanding: the relevant information is no longer confined to written narrative or a single video clip, but distributed across speech, visible behavior, objects, locations, social interactions, and repeated episode-level patterns.

The benchmark contains five public YouTube personal-vlog series, covering 327 videos and roughly 125 hours of audiovisual material. From these sources, KnowMeBenchV2 provides 3,061 evidence-grounded questions with gold answers, source identifiers, task labels, and evaluation metadata. The benchmark is organized around observ-

able evidence. Gold answers are supported by what appears, is said, or repeatedly happens in the source videos, including explicit statements, recurring routines, visual actions, interaction roles, and cross-episode changes. This makes the benchmark suitable for evaluating person understanding as grounded inference over audiovisual traces rather than as open-ended profile generation.

3.1 Task Design

KnowMeBenchV2 is structured as an evidence-scope ladder. T1–T4 move from localized evidence recovery to cross-episode synthesis, while T5 is a contrastive support check for evidence boundaries. Table 2 summarizes the task definitions and answer forms.

T1 focuses on local grounding. These questions require the model to identify a concrete detail from a specified video moment, such as what a person says, holds, does, wears, or encounters. T2 extends this setting to short-range event relations, where the answer depends on connecting neighboring moments within a video rather than copying a single fact. T3 moves beyond single-episode reasoning by asking the model to order evidence instances across videos according to an explicitly stated criterion. These questions test whether the system can preserve provenance and compare events without collapsing them into an unordered set of retrieved clips.

T4 is the main cross-episode synthesis task. It asks questions whose answers are not licensed by any single clip alone. A valid T4 answer must

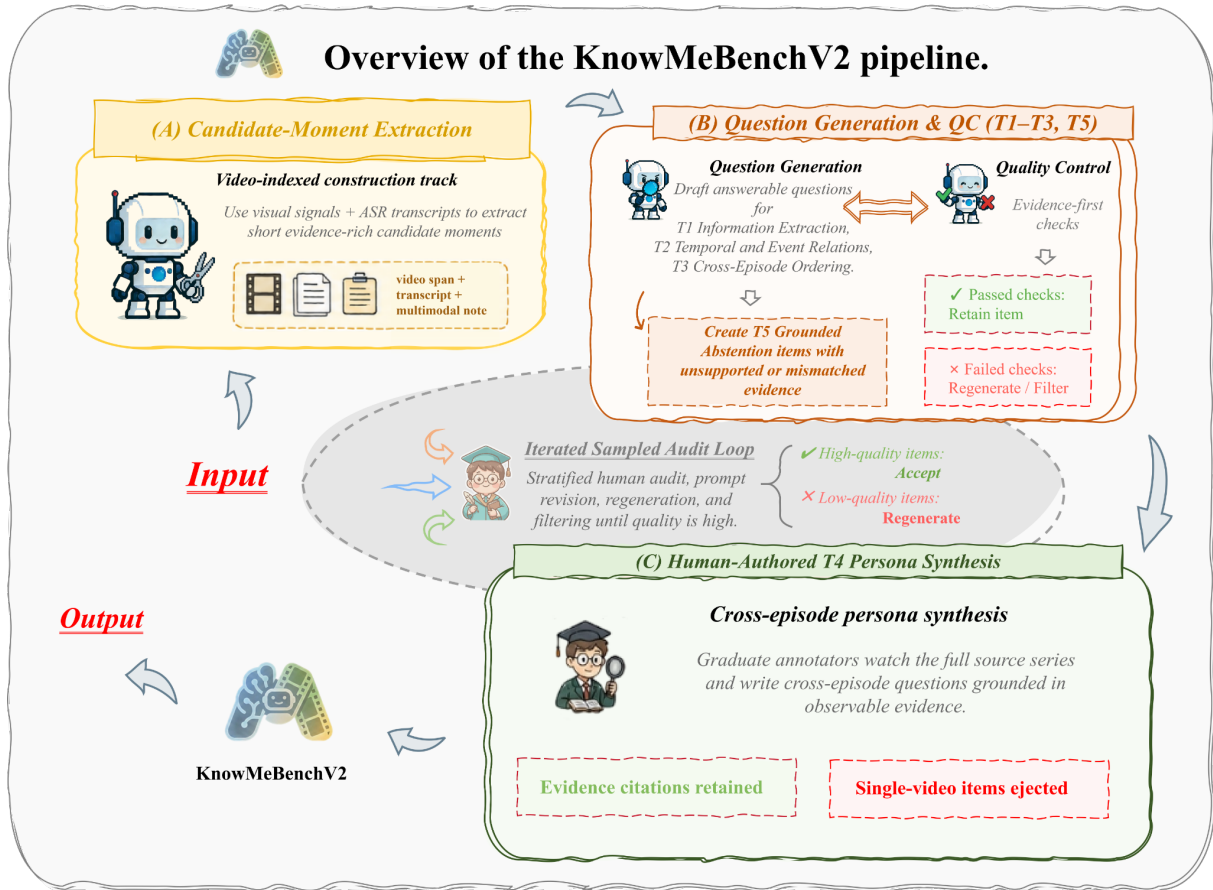


Figure 2: Overview of the KnowMeBenchV2 construction pipeline. (A) Visual signals and automatic speech recognition (ASR) transcripts identify evidence-rich candidate moments. (B) T1–T3 and T5 are drafted from video-indexed evidence and filtered through quality-control gates. (C) T4 is human-authored from full-series viewing and retains cross-episode evidence citations.

be supported by multiple observations from the same source set and must express a recurring pattern at the level of observable behavior, social role, preference, practice, or interaction style. T5 complements the ladder by testing the boundary of support. Its questions deliberately resemble supported person-centric questions in surface form, but combine valid people, settings, objects, or partial events with an unsupported relation, false temporal setting, or missing causal link. The correct answer is ABSTAIN. T5 therefore tests whether a system distinguishes evidential support from surface familiarity, rather than producing a plausible but unsupported person-level explanation.

3.2 Question Construction

KnowMeBenchV2 uses two construction tracks. For T1–T3 and T5, we first extract candidate moments from each source set using visual signals and automatic speech recognition (ASR) transcripts. Candidate moments are short evidence-rich spans

containing person-linked actions, dialogue, objects, settings, or event transitions. Question generation then converts these moments into task-specific items with associated source identifiers and evidence provenance. Quality control checks support, answerability, task fit, abstention polarity, and source traceability. Items that fail these checks are revised, regenerated, or filtered. For T5, construction deliberately combines valid source elements with an unsupported relation or premise, so that familiarity with people, settings, or events is insufficient for answering.

T4 follows a human-authored construction track. Annotators watch the full source series and write cross-episode questions whose answers require evidence distributed across multiple videos. This design aligns the most person-centric task with the natural viewing process: recurring patterns become visible through repeated exposure rather than through isolated retrieval. T4 items are retained only when the answer can be grounded in source

Task	Evidence Scope	Target Capability	Answer Form
T1: Information Extraction	Local clip or moment	Recover grounded facts about a person, action, object, setting, or spoken detail under explicit source constraints.	Short answer
T2: Temporal and Event Relations	Within-video or local episode context	Relate temporally adjacent or causally connected events, including what happened before/after an action or why a local event relation holds.	Short or structured answer
T3: Cross-Episode Ordering	Multiple episodes	Order person-linked events, activities, or evidence instances under a stated criterion, such as recurrence, progression, or observable change.	Ordered list
T4: Cross-Episode Persona Synthesis	Full source set	Synthesize a recurring person-level pattern supported by multiple episodes, such as routines, preferences, roles, interaction habits, or stable behavioral tendencies.	Free-form answer
T5: Grounded Abstention	Unsupported or mismatched evidence	Reject a person-linked premise that is surface-similar to supported questions but not licensed by the source evidence.	ABSTAIN

Table 2: KnowMeBenchV2 task taxonomy. T1–T4 form an evidence-scope ladder from localized extraction to cross-episode synthesis. T5 is evaluated separately as a contrastive support check for grounded abstention.

evidence from more than one episode and when the question targets a stable observable pattern rather than a one-off event. Annotators also record evidence notes used for verification and later audit.

3.3 Human-Authored Synthesis and Quality Audit

Human experts are central to the construction of the most interpretive part of KnowMeBenchV2. Five trained graduate-student annotators each watch one complete source series and write candidate T4 questions whose answers require evidence from multiple separately recorded videos. LLM assistance is limited to wording normalization, duplicate detection, and answer-format standardization; the question, cross-video evidence pattern, and gold answer are human-authored.

Human audit then checks source support, answerability, cross-episode scope, near-duplicate items, and over-interpretation risk. Items are revised or rejected when the evidence supports only a single episode or when the answer requires a claim beyond the observable audiovisual record. For the human reference, annotators answer questions on source series they did not author, using the same source-identifier interface and task-specific scoring protocol as systems. This separates question construction from human answering while keeping both stages grounded in the same audiovisual evidence.

3.4 Scoring Protocol

We use task-specific scoring. T1 uses constrained short-answer matching: exact matching for canonical entities or numbers and normalized equivalence for short textual variants. T2 and T3 are scored for both answer correctness and relation correctness: a response must recover the relevant event relation or ordering criterion, not merely mention plausible evidence. T4 uses semantic grading against the reference answer, emphasizing whether the response captures the recurring pattern and grounds it in source-supported observations.

T5 is strict-scored: full credit is given only to explicit ABSTAIN. We report it separately from the T1–T4 ladder because it measures evidence-boundary recognition rather than additional synthesis depth. Its role is to check whether open-form scores reflect grounded evidence use rather than generic person-level language. This check is important because open-form person-centric questions can otherwise reward plausible but weakly grounded answers. As a no-evidence validity check, Qwen3.5-9B with the evidence route removed scores 1.3/0.0/0.0/0.0 on T1–T4 and 100.0 on T5, indicating that the supported tasks cannot be solved from question priors alone while T5 preserves its intended unsupported-premise signal. On semantic scoring, a 100-record blind audit by three NLP researchers gives 88.0% exact agreement and Cohen’s $\kappa=0.78$ against the GPT-5.2 judge, with full disagreement analysis in Appendix E. Full scoring prompts, audit procedures, agreement analysis, and task examples are

provided in the appendix.

4 Experiments

The experiments evaluate how current systems use audiovisual evidence at different scopes. We report end-to-end long-video systems and a fixed-answerer evidence diagnostic. The first view characterizes current system-level behavior on KnowMeBenchV2; the second view isolates how performance changes when the same answerer receives different evidence representations.

4.1 Experimental Setup

Systems. We evaluate representative long-video and multimodal systems, including LongVideo-R1 (Qiu et al., 2026), M3-Agent (Long et al., 2026), VideoRAG (Ren et al., 2026), and MA-LMM (He et al., 2024). These systems differ in their video access mechanisms, retrieval modules, temporal reasoning procedures, and memory designs. Their results characterize current long-video behavior across localized and cross-episode tasks.

Fixed-answerer evidence diagnostic. We additionally run a fixed-answerer diagnostic with Qwen3.5-9B. Qwen3.5-9B is a multimodal-capable answerer; in this diagnostic, we hold the answer model fixed and vary the evidence representation supplied at inference time. ASR-Retrieval provides retrieved speech-transcript chunks, while KnowMe provides routed textual evidence summaries constructed from audiovisual source evidence. Both routes use the same task-specific pre-tokenization character cap. The answerer receives textual context in both conditions, so this diagnostic isolates evidence construction and organization rather than native video-token processing.

KnowMe first builds an offline hierarchy of clip, segment, video-level persona, and series-level trajectory memories. At inference time, the answerer does not receive the full hierarchy. Each question retrieves a task-specific subset: local tasks primarily use clip and segment evidence, while cross-episode tasks additionally route through persona and trajectory memories before returning supporting lower-level evidence.

Metrics. We report T1–T5. T1–T4 form the main person-understanding ladder and are averaged as Overall. We also report Core, the average of T3 and T4, to emphasize cross-episode comparison and synthesis. T5 is reported separately because

it measures evidence-boundary recognition: the correct behavior is to abstain when a person-linked premise is unsupported by the source evidence.

4.2 Main Results

Table 3 shows a split between local video grounding and cross-episode evidence organization. Human viewers remain far ahead, reaching 85.0 Overall and 80.0 Core, which shows that the questions are recoverable from the audiovisual source evidence while leaving a substantial gap for current systems.

Among non-human systems, LongVideo-R1 performs best on the localized tasks T1 and T2, with scores of 44.0 and 52.0. This indicates that native video access is useful when the answer is close to a specific observed moment. In contrast, Qwen3.5-9B+KnowMe performs best on cross-episode synthesis, reaching 64.9 on T4 and 57.0 Core. The contrast shows that local video grounding and cross-episode evidence organization are distinct capabilities. It does not establish textualized evidence as universally superior to native-video systems: LongVideo-R1 is one meaningful native-video reference, and stronger native-video ceilings remain to be tested.

T5 provides a separate check on evidence boundaries. Because T5 items are surface-similar to supported person-linked questions but unsupported by source evidence, high T5 performance indicates that a system can reject plausible but unlicensed claims. The two Qwen3.5-9B diagnostic routes score 97.0 and 100.0 on T5, supporting the interpretation that their open-form scores are not produced solely by generic person-level language. We therefore report T5 outside Overall and Core: it supports the reliability of open-form evaluation, but does not measure additional synthesis depth.

4.3 Attribution Diagnostics

Figure 3 separates transcript volume from evidence organization under the same Qwen3.5-9B answerer. ASR-Flat-32K and ASR-Retrieval remain close on Overall, while KnowMe nearly doubles the score. The largest difference appears on T4: ASR-Retrieval reaches 2.7, KnowMe without visual fields reaches 51.9, and full KnowMe reaches 64.9. The 2.7→51.9 gain (+49.2) is associated with hierarchical summarization, structured routing, and cross-episode organization of ASR-derived evidence; adding vision-language-model (VLM)-extracted visual fields contributes a further

System	T1 Extract.	T2 Temporal	T3 Ordering	T4 Synthesis	Overall T1–T4	Core T3–T4	T5 Abstain
Qwen3.5-9B ASR-Retrieval	26.7	15.0	42.9	2.7	21.8	22.8	97.0
Qwen3.5-9B + KnowMe	34.0	47.0	49.0	64.9	48.7	57.0	100.0
LongVideo-R1	44.0	52.0	32.7	56.8	46.4	44.8	76.0
M3-Agent	17.3	39.0	49.0	18.9	31.0	34.0	45.0
VideoRAG	3.0	17.8	24.4	2.7	12.0	13.6	67.0
MA-LMM	4.7	29.0	22.4	–	–	–	12.0
Human reference	92.0	88.0	82.0	78.0	85.0	80.0	93.0

Table 3: Main results on KnowMeBenchV2. Overall is the average over T1–T4. Core is the average over the cross-episode tasks T3–T4. T5 is reported separately because abstention evaluates evidence-boundary recognition rather than synthesis depth. Bold numbers mark the strongest non-human result for each T1–T4 column and aggregate.

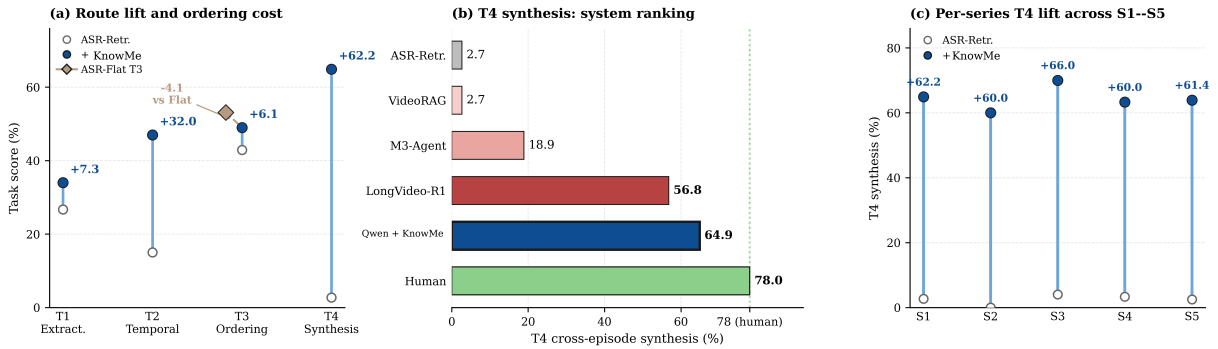


Figure 3: **Attribution diagnostics for the fixed-answer route comparison.** (a) Route lift across T1–T4 under the same Qwen3.5-9B answerer. Gray circles denote ASR-Retrieval, blue circles denote KnowMe, and the brown diamond marks the ASR-Flat-32K boundary condition on T3. KnowMe improves T1, T2, and especially T4, while ASR-Flat-32K remains stronger on T3, revealing an ordering cost of hierarchical compression. (b) System ranking on T4 cross-episode synthesis. KnowMe reaches 64.9, exceeding all evaluated systems and approaching the human reference of 78.0. (c) Per-series T4 lift across the five source series (S1–S5). The synthesis gain is consistent across all source sets, indicating that the route effect is not driven by a single series.

+13.0 points within the same hierarchy. The full 62.2-point gap should therefore not be attributed entirely to vision.

Panel (a) also exposes a compression tradeoff. ASR-Flat-32K exceeds KnowMe on T3, 53.1 versus 49.0, even though it fails on T4; M3-Agent also reaches 49.0, tying KnowMe on this task. Ordering can depend on local discriminative cues preserved in a long transcript, while synthesis benefits from higher-level organization. We therefore treat T3 as a current limitation of hierarchical compression rather than evidence of a universal KnowMe advantage.

Two controls test whether the T4 route effect is specific to one model configuration. On a preselected 37-question T4 subset from one 50-episode source series, KnowMe is directionally better than ASR-Retrieval with Qwen3.5-9B (21.6→40.5), GLM-5.2 (27.0→32.4), and GPT-5.2 (70.3→73.0) as final answerers. The effect magnitude varies

substantially, and the GPT-5.2 comparison is descriptive because GPT-5.2 also serves as the semantic judge. In a separate cross-summarizer control with Qwen3.5-9B fixed as audiovisual extractor and final answerer, independently generated Qwen, GLM-5.2, and GPT-5.2 hierarchies score 40.5, 43.2, and 67.6, respectively, versus 21.6 for ASR-Retrieval. These subset results reduce the concern that the gain is confined to one answerer or solely to Qwen consuming its own summaries; detailed raw counts and protocols are in Appendix G.1.

Panel (c) further shows that the T4 gain is stable across all five source series, ranging from +60.0 to +66.0 points over ASR-Retrieval. Additional route ablations confirm this interpretation. Removing the trajectory/theme layer reduces T4 from 64.9 to 43.8, while leaving lower-level tasks comparatively less affected. KnowMe also uses a smaller final answer-time context than transcript-heavy controls: 1.8K–2.8K tokens including prompt and routed ev-

idence versus about 8K–12K for the task-matched ASR-Retrieval route and about 28K for ASR-Flat-32K (Appendix G.2). Full component ablations, routing budgets, and qualitative examples are reported in Appendix G.

5 Conclusion

We introduced KnowMeBenchV2, an evidence-grounded benchmark for person-centric long-video understanding over recurring subjects in multi-episode audiovisual sources. The benchmark evaluates a ladder of evidence scopes, from localized extraction to cross-episode synthesis, and uses grounded abstention as a contrastive check on evidence-boundary recognition.

The experiments show that current systems fail in task-dependent ways. Native long-video systems are strongest on local grounding, while structured textual representations constructed from audiovisual evidence are effective for cross-episode synthesis. Under a fixed Qwen3.5-9B answerer, replacing flat ASR retrieval with the KnowMe evidence route raises T4 from 2.7 to 64.9, approaching the human reference of 78.0. At the same time, ordering tasks reveal a compression cost: summaries help expose recurring patterns but can discard rank-sensitive details.

KnowMeBenchV2 reframes recurring-person video understanding as an evidence-construction problem over audiovisual traces. Progress on this setting requires systems that preserve local visual grounding while organizing evidence across episodes into representations suitable for comparison, synthesis, and abstention.

Limitations

Source breadth and language. The source collections are drawn from five public YouTube personal-vlog series; source speech is primarily English, with a non-English subset, while benchmark questions and gold answers are English throughout. Results on a balanced multilingual or low-resource setting may differ.

Textualized audiovisual evidence and native-video ceilings. The fixed-answerer diagnostic does not feed raw video tokens to the final answerer. It first constructs textual representations from audiovisual evidence and then performs question-conditioned routing. These experiments therefore support claims about audiovisual evidence construction and cross-episode organization, not a gen-

eral claim that textualized evidence is superior to native-video processing. LongVideo-R1 provides one meaningful native-video reference, but stronger native-video models may narrow or reverse the observed gap.

Hierarchical compression. The same hierarchy that helps T4 synthesis can remove local discriminative cues needed for ordering. ASR-Flat-32K outperforms KnowMe on T3 (53.1 versus 49.0), and KnowMe ties M3-Agent at 49.0. Future systems should combine global abstraction with mechanisms that preserve rank-sensitive local evidence.

Ethical Considerations

KnowMeBenchV2 is intended for evidence-grounded long-video research. The task definitions restrict behavioral claims to observable source evidence and do not support surveillance, identity inference, protected-attribute inference, psychological profiling, or claims about sensitive attributes such as health, religion, sexuality, or political identity. Target subjects are recurring people who are central to the public source series; incidental bystanders are not benchmark subjects and are not used for persona claims. Cross-episode synthesis items ask for recurring evidence patterns within a source series, not stable psychological labels or protected-attribute inferences. The source videos are public personal vlogs hosted on YouTube and are not redistributed by us. We have not obtained blanket redistribution permission for benchmark annotations or source-derived artifacts and are contacting the original creators to clarify those permissions. Accordingly, this camera-ready version makes no blanket public-release commitment and assigns no blanket data license to those materials; any future availability must follow the permissions obtained and applicable platform, privacy, and rights constraints.

References

- Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2025. [InfiniBench: A benchmark for large multi-modal models in long-form movies and TV shows](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19485–19512, Suzhou, China. Association for Computational Linguistics.
- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristóbal Eyz-

- guirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. 2024. [HourVideo: 1-hour video-language understanding](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Zhisheng Chen, Tingyu Wu, Zijie Zhou, Zhengwei Xie, Jinhan Li, Ziyang Weng, Liang Lin, Jingwei Song, Zikai Xiao, and Yingwei Zhang. 2026. [PolarMem: A training-free polarized latent graph memory for verifiable vision-language models](#). *arXiv preprint arXiv:2602.00415*.
- Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. 2023. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24049–24059.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. [Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.
- Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of CVPR*.
- Yubo Hou, Zhisheng Chen, Tao Wan, and Zengchang Qin. 2026a. [FlashMem: Distilling intrinsic latent memory via computation reuse](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 4687–4705, San Diego, California, United States. Association for Computational Linguistics.
- Yubo Hou, Jingwei Song, Hongbo Zhang, Zhisheng Chen, Bang Xiao, Tao Wan, and Zengchang Qin. 2026b. [PersonaTree: Structured lifecycle memory for person understanding in llm agents](#). *arXiv preprint arXiv:2606.04780*.
- Sen Hu, Zhiyu Zhang, Yuxiang Wei, Xueran Han, Zhenheng Tang, Ronghao Chen, and Huacan Wang. 2026. [CloneMem: Benchmarking long-term memory for AI clones](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33571–33602, San Diego, California, United States. Association for Computational Linguistics.
- Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10854–10863.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of EMNLP*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. MV-Bench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of CVPR*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*.
- Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. 2026. [Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory](#). In *International Conference on Learning Representations*.
- Romy Luo, Zihui (Sherry) Xue, Alex Dimakis, and Kristen Grauman. 2025. [When thinking drifts: Evidential grounding for robust video reasoning](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhui Chen. 2026. [VideoEval-Pro: Robust and realistic long video understanding evaluation](#). *Transactions on Machine Learning Research*.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of ACL*.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

- Yiran Meng, Junhong Ye, Wei Zhou, Guanghui Yue, Xudong Mao, Ruomei Wang, and Baoquan Zhao. 2025. [VideoForest: Person-anchored hierarchical reasoning for cross-video question answering](#). In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 836–845.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. [Memgpt: Towards llms as operating systems](#). Preprint, arXiv:2310.08560.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*.
- Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. 2026. [LongVideo-R1: Smart navigation for low-cost long video understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40505–40515.
- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2026. [VideoRAG: Retrieval-augmented generation with extreme long-context videos](#). In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pages 2390–2401.
- Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. [MovieChat: From dense token to sparse memory for long video understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.
- Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. 2024. [Selective “selective prediction”: Reducing unnecessary abstention in vision-language reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12935–12948, Bangkok, Thailand. Association for Computational Linguistics.
- Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2026. [Beyond dialogue time: Temporal semantic memory for personalized LLM agents](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29935–29951, San Diego, California, United States. Association for Computational Linguistics.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefel hagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. [MovieQA: Understanding stories in movies through question-answering](#). In *Proceedings of CVPR*.
- Weihan Wang, Zehai He, Wen yi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. [LVBench: An extreme long video understanding benchmark](#). *arXiv preprint arXiv:2406.08035*.
- Zhaoyang Wei, Guoliang Wang, Guohua Gao, Yanchao Hao, Mingda Li, Wenchao Ding, Xi Chen, Shizhu He, and Xuehui Yu. 2026. [Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10584–10592.
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. [LongMemEval: Benchmarking chat assistants on long-term interactive memory](#). In *International Conference on Learning Representations*.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. [LongVideoBench: A benchmark for long-context interleaved video-language understanding](#). In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.
- Tingyu Wu, Zhisheng Chen, Ziyan Weng, Shuhe Wang, Shuo Zhang, Sen Hu, Silin Wu, Qizhen Lan, Huacan Wang, and Ronghao Chen. 2026. [KnowMe-Bench: Benchmarking person understanding for lifelong digital companions](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30214–30238, San Diego, California, United States. Association for Computational Linguistics.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. 2019. [ActivityNet-QA: A dataset for understanding complex web videos via question answering](#). In *Proceedings of AAAI*.
- Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. 2025. [VideoExplorer: Think with videos for agentic long-video understanding](#). *arXiv preprint arXiv:2506.10821*.
- Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. 2025. [SeriesBench: A benchmark for narrative-driven drama series understanding](#). In *Proceedings of CVPR*.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [MemoryBank: Enhancing large language models with long-term memory](#). In *Proceedings of AAAI*.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. [MLVU: Benchmarking multi-task long video](#)

[understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701.

Ting Zhou, Daoyuan Chen, Qirui Jiao, Bolin Ding, Yaliang Li, and Ying Shen. 2024. [HumanVBench: Exploring human-centric video understanding capabilities of MLLMs with synthetic benchmark data](#). *arXiv preprint arXiv:2412.17574*.

Statistic	S1	S2	S3	S4	S5	Total	Scoring
Source videos	50	55	91	75	56	327	–
Main questions	439	513	859	684	566	3,061	–
Video-indexed	402	488	809	654	480	2,833	–
Series synthesis	37	25	50	30	86	228	–
T1 Information extr.	150	165	273	225	168	981	constrained
T2 Temp.-event rel.	100	110	182	150	112	654	strict
T3 Ordering	49	49	49	49	49	245	semantic
T4 Persona synth.	37	25	50	30	86	228	semantic
T5 Grounded abst.	100	110	182	150	112	654	strict

Table 4: Benchmark construction statistics. S1–S5 denote the five source series. The T4 row reports the full 228-item constructed pool; all reported cross-system T4 comparisons use the separate 185-item complete-record evaluation subset described below. T3 is balanced across series, while T4 is human-authored with a series-dependent yield.

A Dataset Scope and Availability

KnowMeBenchV2 uses five public person-centric YouTube personal-vlog series: 327 source videos across five playlists, totaling roughly 125 hours. Source speech is primarily English, with a subset of videos containing additional non-English speech; benchmark questions and gold answers are in English throughout. Raw videos are not redistributed. We have not obtained blanket permission from the original creators to redistribute benchmark annotations or source-derived artifacts such as ASR segments, frame timestamps, visual-extraction fields, generated memories, or retrieval outputs. We are contacting the creators to clarify those permissions and therefore make no blanket public-release commitment or blanket data-license claim for these materials in the camera-ready version.

Internal provenance records retain source identifiers, platform links, durations, language metadata, access dates, evidence anchors, and availability status. These records support construction and evaluation. Any future redistribution of benchmark or source-derived materials must follow the permissions obtained and applicable platform, privacy, and rights constraints.

Table 4 reports the full constructed question pool. T3 uses its full balanced 245-item pool (49 items per series) for the reported cross-system comparisons. T4 contains 228 constructed items, but every reported cross-system T4 comparison uses the same 185-question complete-record subset for which each evaluated system has both a prediction and a semantic-judgment record. The T4 subset criterion is record completeness and is independent of answer correctness. In particular, KnowMe has 120

correct predictions on this subset, giving $120/185 = 64.9$. The identical count of 120 correct predictions on T3 ($120/245 = 49.0$) is coincidental.

ASR-Flat-32K input construction. For single-video tasks (T1–T3 and T5), the ASR-Flat-32K boundary feeds the answerer the full transcript of the target video, capped at 110K characters ($\approx 28K$ tokens) by head-50% + tail-50% truncation. For T4, per-video ASR is concatenated in episode order across the corresponding series source set and truncated front-to-back to the same character cap, so only a small fraction of the source set reaches the answerer’s context window. Both modes use a 32K context window, no visual input, answerer temperature 0.1, and a 2,048-token output cap.

B Task Definitions and Examples

T1: Information Extraction. T1 asks for localized facts, entities, attributes, numerical details, or time-anchored observations directly supported by the video or transcript. It uses constrained short-answer matching: exact matching for canonical entities or numbers and normalized equivalence for short textual variants.

T2: Temporal and Event Relations. T2 covers before/after relations, local event sequences, short-window reactions, and event-linked social cues within a video. It is strict-scored.

T3: Cross-Episode Ordering. T3 asks the system to rank a small set of people, events, or items from the same source series along a stated criterion such as severity, urgency, escalation, risk, cost, recurrence, or observable change. Each item lists 3–5 candidates and is judged for semantic equivalence to the gold ordering rather than exact string match.

T4: Cross-Episode Persona Synthesis. T4 integrates evidence across multiple videos to identify recurring stated preferences, roles, practices, routines, or observable behavior patterns within the associated source collection. Each item is associated with the series-specific source set and is semantically scored. In the evaluated T4 pool, ABSTAIN is scored incorrect because retained T4 items are answerable from the source set.

T5: Grounded Abstention. T5 is a contrastive support check for open-form person-centric evaluation. Items are surface-similar to supported questions but combine valid people, settings, objects, or partial events with an unsupported relation, false

temporal setting, or missing causal link. The gold answer is the fixed token ABSTAIN. T5 is strict-scored and reported separately from T1–T4.

Representative examples. Construction records include item-level examples with source identifiers, timestamps, evidence notes, and answers. T1 examples ask for a concrete local detail, T2 examples ask about an event relation, T3 examples ask for a ranked list under a stated criterion, T4 examples ask for a recurring person-level pattern supported by multiple videos, and T5 examples contain an unsupported premise while retaining familiar entities or settings.

C Annotation and Human Materials

Gold answers were constructed with access to full video context and localized evidence. Per-series construction records contain video-indexed main-task records and cross-episode synthesis records, each with question id, task type, video id (or video-set id for T4), localized time range (or per-video evidence anchors for T4), question, answer, and annotator metadata. These records support provenance for gold-answer construction but are not used as model inputs.

The cross-episode synthesis construction material comes from five trained graduate-student annotators, with one annotator assigned to each source series. After calibration, each annotator watched the full source set for the assigned series and wrote candidate questions whose answers require evidence from multiple separately recorded videos. LLM assistance was limited to wording normalization, duplicate detection, and answer-format standardization; the question, cited cross-video evidence pattern, and gold answer were human-authored. Human evidence checks merged candidates, removed near-duplicates, checked source support, answerability, cross-episode scope, and over-interpretation risk, and rejected single-video items, retaining 228 final items with provenance links to supporting source videos.

Construction-stage agreement on a separate 60-item stratified double-annotated subset (12 items per task) gives per-task Cohen’s κ of 0.78 (T1), 0.71 (T2), 0.66 (T3), 0.61 (T4), and 0.74 (T5), with aggregate $\kappa = 0.70$. T3 and T4 show lower agreement than the more constrained-scoring tasks, consistent with semantic-scoring tasks being harder to standardize at the gold level.

Internal construction label	Paper-level task
code-T1	T1 Information Extraction
code-T2	T5 Grounded Abstention
code-T3	T2 Temporal and Event Relations
code-T4	Per-video ordering draft $\rightarrow$ T3
Human-authored synthesis	T4 Cross-Episode Persona Synthesis

Table 5: Mapping between historical construction-pipeline labels and the T1–T5 task names used throughout the paper. The mapping is documented explicitly because the internal labels predate the final paper taxonomy.

D Question Construction Prompts and QC Gates

The construction pipeline uses prompt cards rather than unconstrained generation. Table 5 records the internal-to-paper task-label mapping used by the prompt cards below. Each card defines the role, input fields, output schema, rejection gate, and hard constraints. This appendix summarizes the contracts needed to reproduce the paper-level items.

Candidate-moment extraction. The candidate-moment prompt watches one full video and returns 5–15 evidence-rich windows with `video_id`, start time, end time, and a short factual reason. Candidate moments are selected for clear speech content, visible action changes, environmental changes, reactions, anomalies, contrasts, or causal structure. The prompt requires JSON-only output and forbids invented events, motives, or background facts.

Supported video-indexed item generation. For T1 and T2, the item-generation prompt conditions on candidate moments and ASR/visual notes. It requires a uniquely answerable question, a concise answer, task label, source video id, timestamp span, and evidence note. The QC prompt rejects items whose answer is not supported by the cited span, whose wording leaks the answer, whose answer is not unique, or whose task label does not match the intended capability.

Cross-episode ordering construction. T3 construction aggregates candidate events within the same source series and asks for a ranked list under an explicitly stated observable criterion. The prompt requires discrete evidence instances, a non-trivial ranking dimension, and evidence closure: the cited evidence must be sufficient to recover both the candidates and the ordering rationale.

Grounded abstention construction. T5 construction deliberately combines true fragments into a false or unsupported premise. Entities, people, settings, or partial events are valid, but the relation is invalid. Trap strategies include entity swapping, spatiotemporal distortion, and false causality via unrelated true anchors. The answer must be exactly ABSTAIN.

Quality-control gates. The shared QC pass checks support, answerability, task fit, abstention polarity, observable grounding, provenance completeness, and over-interpretation risk. Items that fail are revised, regenerated, or filtered. T4 additionally requires multi-episode support and human evidence notes; T5 additionally requires that the unsupported premise be surface-similar to supported items while still unanswerable from the source.

A. Candidate-Moment Extraction

Role	Identify candidate moments from one full source video that are suitable for downstream question generation.
Inputs	Uploaded source video and source-level video_id.
Output Contract	A JSON list of 5–15 candidate moments. Each record contains video_id, start_time, end_time, and a short factual reason.
Gate	If no valid candidate moment can be generated, return an empty list. Do not output fields beyond the specified schema.

Hard Constraints.

- **Source faithfulness:** use only content actually present in the video; do not invent events, motives, or background facts.
- **Moment quality:** prefer concrete and visually/verbal clear moments.
- **Selection criteria:** select moments with clear speech content, visible action change, environmental change, reaction, anomaly, contrast, causal structure, or evidence likely to support a unique-answer or stable-ordering question.
- **Reason format:** keep the reason short and factual.
- **Schema-only output:** return JSON only.

Skeleton.

```
Input: uploaded video + video_id.  
Operation: extract 5--15 evidence-rich candidate moments.  
Output(JSON): [{video_id, start_time, end_time, reason}, ...]
```

B. Per-Video Drafting: code-T1 / code-T2 / code-T3

Role	Generate per-video benchmark QA drafts from the uploaded video and the provided candidate moments.
Inputs	Uploaded source video and candidate_moments_json.
Output Contract	A JSON list containing 3 code-T1 items, 2 code-T2 items, and 2 code-T3 items. Each record contains question_id, task_type, question, answer, video_id, and time_range.
Gate	If no valid item can be generated, return an empty list. Do not output fields beyond the specified schema. Retention is decided by downstream evidence and audit checks.

Task Mapping.

- **code-T1**: fact extraction with a unique verifiable answer; this corresponds to paper-T1 information extraction.
- **code-T2**: adversarial abstention by combining real elements into a false scenario; this corresponds to paper-T5 grounded abstention.
- **code-T3**: temporal reasoning such as duration, order, before/after relations, or event sequence; this corresponds to paper-T2 temporal/event relations.

Hard Constraints.

- **Source faithfulness**: use only information actually present in the video; do not invent unseen events or unsupported causal relations.
- **code-T1 uniqueness**: answers must be unique and factual.
- **code-T2 abstention**: answers must be exactly ABSTAIN.
- **code-T3 temporal dependence**: questions must depend on temporal reasoning, not only surface lookup.
- **Quality control**: prefer clear, high-confidence questions and avoid duplicates.
- **Schema-only output**: return JSON only.

Skeleton.

```
Inputs: uploaded video + candidate_moments_json.  
Generate: 3 code-T1 + 2 code-T2 + 2 code-T3 draft items.  
Output(JSON): [{question_id, task_type, question, answer, video_id, time_range}, ...]  
Constraint: code-T2 answer = ABSTAIN.
```

C. Per-Video Drafting: code-T4 Logical Ordering

Role	Generate one logical-ordering draft item from the uploaded video and the provided candidate moments.
Inputs	Uploaded source video and candidate_moments_json.
Output Contract	A JSON list containing one code-T4 item. Each record contains question_id, task_type, question, answer, video_id, and time_range. The answer is a comma-separated string representing the ordered events.
Gate	If no valid item can be generated, return an empty list. Do not output fields beyond the specified schema. These per-video drafts feed the cross-episode aggregation that produces the paper-T3 ordering probe.

Hard Constraints.

- **Source faithfulness:** use only events actually present in the video.
- **Comparable events:** select 3–4 comparable events.
- **Non-temporal ordering:** the ordering must not be based on time.
- **Explicit semantic dimension:** rank events by a stated dimension such as risk, physical effort, environmental harshness, emotional intensity, or equipment difficulty.
- **Objectivity:** the ranking dimension should be as objective as possible; avoid highly subjective or ambiguous criteria.
- **Answer format:** return a comma-separated string representing the ordered events.
- **Schema-only output:** return JSON only.

Skeleton.

```
Inputs: uploaded video + candidate_moments_json.  
Generate: 1 code-T4 logical-ordering draft item.  
Choose: 3--4 comparable events + explicit non-temporal semantic dimension.  
Output(JSON): [{question_id, task_type, question, answer, video_id, time_range}]  
Use: per-video code-T4 draft -> cross-episode aggregation -> paper-T3 ordering probe.
```

Subset	N	Correct	Acc.	Conf.
T3 ordering	245	120	49.0	0.84
T4 synthesis	185	120	64.9	0.78
Audit subset	100	88	88.0	–

Table 6: Semantic scorer validation summary for the reported Qwen3.5-9B + KnowMe semantic-scored subsets and the blind human audit subset.

E Scoring and Semantic-Judge Audit

T1, T2, and T5 use strict or constrained matching. T3 and T4 use semantic scoring because valid answers are often paraphrastic: T3 admits any ordering that preserves the gold ranking on the named criterion, and T4 admits any synthesis that names the same recurring pattern using source-equivalent observable evidence. A fixed GPT-5.2 judge at temperature 0.0 takes the question, gold answer, and prediction and returns JSON fields for equivalence, confidence, and rationale. Core claim, stance, implication, and yes/no polarity must match; factual contradictions are marked incorrect.

Table 6 summarizes the semantic-scorer validation. For the 100-record audit subset, three blind NLP researchers label each model prediction by majority vote. The GPT-5.2 judge matches the human majority with 88.0% exact agreement and Cohen’s $\kappa = 0.78$, against a mean pairwise human ceiling of $\kappa = 0.82$. The record-level Pearson correlation between judge and human labels is $r = 0.86$ ($p < 0.001$). A third-party Claude Opus 4.7 cross-judge audit on the same records preserves the ASR-Flat-32K versus KnowMe ordering on T3 and T4, ruling out a scorer-family-specific artifact.

No-evidence validity check. With no evidence route attached, the Qwen3.5-9B answerer scores 1.3/0.0/0.0/0.0 on T1–T4 and 100.0 on T5. This validates two intended properties of the benchmark: supported questions require source evidence rather than question priors, and T5 retains its unsupported-premise signal even without evidence.

F Fixed-Answerer Evidence Routes

F.1 Protocol

KnowMe is the reference evidence route used for the fixed-answerer diagnostic. ASR-Retrieval and KnowMe share the same question, source identifiers, answerer, prompt family, and scoring protocol; only the context-construction route differs. The diagnostic compares evidence representations

supplied to the same multimodal-capable Qwen3.5-9B answerer, not the native modality capacity of the backbone.

KnowMe segments each video into 30-second clip units with ASR text, sampled keyframes, and vision-language-model-derived visual fields. It then groups clips into segment memories, summarizes video-level persona cues, and builds series-level trajectory/theme memories. At query time, T1 retrieves local clip and segment evidence, T2 emphasizes event sequences, T3 traverses persona-level cues plus cross-episode segment evidence for the ordering criterion, T4 starts from trajectory and theme nodes before retrieving supporting evidence, and T5 reuses the local evidence route while checking premise support.

F.2 Engineering Stack

ASR is stored as transcript JSON with segment-level timing. Video is partitioned into 30-second non-overlapping clips with three keyframes per clip at offsets 0.2, 0.5, and 0.8. Visual extraction uses Qwen3.5-9B as the primary extractor and MiniCPM-V-2.6-int4 as fallback, producing scene, visible text, people, objects, actions, alignment cues, uncertainty, and summary fields. Segments group four clips with stride two. Segment, persona, and trajectory summaries are generated by Qwen3.5-9B at temperature 0.0 under schema-constrained output.

Embeddings come from BGE-Multilingual-Gemma2 via SentenceTransformers with normalized vectors. Indexing uses FAISS IndexFlatIP for dense retrieval and BM25Okapi for sparse retrieval, fused into a hybrid BM25+dense pipeline that feeds task-aware hierarchical retrieval over clip, segment, persona, and theme memories. The answerer is Qwen3.5-9B at temperature 0.1 with evidence-first formatting and explicit ABSTAIN support.

F.3 Route–Task Compatibility

A context-construction route is task-compatible with a task when its input-construction policy can in principle deliver evidence from anywhere in the source set at the coverage required by that task. When a route is not task-compatible, its cell is reported as “–” and excluded from paired diagnostic aggregates for that task. This convention applies to MA-LMM’s T4 cell because its input regime does not support the series-level T4 setting.

ASR-Flat-32K is treated as a boundary transcript condition rather than the primary task-matched

Route	Evidence source	Retrieval	Routing budget	Cross-video access
ASR-Flat-32K	Full ASR, truncated	None	110K chars	Truncated window
ASR-Retrieval	Retrieved ASR chunks	Hybrid	Task-matched cap	Series retrieval (T4)
KnowMe full	Clip-to-trajectory memories	Hybrid + hierarchy	Same cap	Persona/theme routing

Table 7: Evidence-source design for attribution controls. “Hybrid” denotes the BM25+dense retrieval pipeline described in Appendix F.2. ASR-Retrieval and KnowMe use the same task-specific pre-tokenization character cap; ASR-Flat-32K is a separate boundary condition. Exact routing limits and resulting context lengths are reported in Appendix G.2.

Route	T1	T2	T3	T4	Overall	T5
ASR-Flat-32K	28.0	17.0	53.1	0.0	24.5	97.0
ASR-Retrieval	26.7	15.0	42.9	2.7	21.8	97.0
KnowMe full	34.0	47.0	49.0	64.9	48.7	100.0

Table 8: Matched-answerer attribution results under Qwen3.5-9B. The table separates transcript volume from organized audiovisual evidence.

route. It can be informative on local and ordering tasks, where relevant evidence may fall inside the retained transcript window, but it is not a sufficient route for full-series persona synthesis because T4 evidence patterns can be distributed across the source set beyond the truncation window.

G Task-Matched Routing Attribution and Route Ablations

Tables 7 and 8 specify the route comparison, while Table 9 isolates hierarchy components. The attribution controls show that transcript volume and evidence organization have different effects. ASR-Flat-32K is strongest on T3 but fails on T4, indicating that long transcripts can preserve ordering-relevant details while still failing to expose recurring cross-episode patterns. Removing trajectory/theme memory reduces T4 substantially, showing that the synthesis gain depends on cross-episode aggregation. The KnowMe w/o visual fields ablation retains segment, persona, and trajectory/theme summaries built from ASR-derived clip evidence while removing VLM-extracted scene, optical-character-recognition (OCR), action, object, and interaction fields. On T4, ASR-Retrieval 2.7→KnowMe w/o visual fields 51.9 (+49.2) isolates the combined contribution associated with hierarchical summarization, structured routing, and cross-episode organization of ASR-derived evi-

Setting	T2	T3	T4
Full KnowMe	47.0	49.0	64.9
w/o trajectory/theme	45.6	47.6	43.8
w/o segment summaries	42.9	46.2	47.0
Clip-only	41.6	44.9	33.0
KnowMe w/o visual fields	41.3	44.5	51.9

Table 9: Component ablations under the Qwen3.5-9B answerer and summarizer. “KnowMe w/o visual fields” retains the full segment–persona–trajectory hierarchy but derives clip-level evidence from ASR only, removing VLM-extracted visual fields. T1 and T5 are omitted because they are less sensitive to the hierarchy components.

Final answerer	ASR-Retrieval	KnowMe
Qwen3.5-9B	8/37 (21.6)	15/37 (40.5)
GLM-5.2	10/37 (27.0)	12/37 (32.4)
GPT-5.2	26/37 (70.3)	27/37 (73.0)

Table 10: Cross-answerer control on the fixed 37-question T4 subset. Qwen3.5-9B generates the audiovisual evidence hierarchy, which is held fixed across final answerers. The route effect is directionally positive for all three answerers, but its magnitude varies. The GPT-5.2 comparison is descriptive because GPT-5.2 also serves as the semantic judge.

dence; 51.9→64.9 (+13.0) is the additional contribution of VLM-extracted visual evidence within the same hierarchy and answerer setting.

G.1 Cross-Answerer and Cross-Summarizer Controls

Both supplementary controls use one source series fixed before the runs, containing 50 episodes and all 37 T4 questions from that series. The subset was not selected based on model correctness, and the same semantic-judging protocol is used throughout.

Table 10 shows that the route effect is not confined to a single final answerer, although the 37-item subset does not support a general capacity-scaling conclusion or significance claim. Table 11 addresses summarizer–answerer circularity: independently generated non-Qwen hierarchies also outperform ASR-Retrieval and can exceed the Qwen self-summary condition. These results reduce the concern that self-alignment is the sole explanation for the route gain, but do not establish a definitive hierarchy-generator ranking or independence from the fixed audiovisual extractor.

G.2 Token Budget and Memory Corpus

KnowMe trades a single long-context transcript pass for a structured, retrieved-evidence pass. The

Hierarchy generator	Correct	Accuracy
ASR-Retrieval baseline	8/37	21.6
Qwen3.5-9B self-summary	15/37	40.5
GLM-5.2 hierarchy	16/37	43.2
GPT-5.2 hierarchy	25/37	67.6

Table 11: Cross-summarizer control on the same T4 subset. Qwen3.5-9B is fixed as the audiovisual extractor and final answerer, while Qwen3.5-9B, GLM-5.2, and GPT-5.2 independently generate all segment, persona, and trajectory/theme hierarchy levels from the same Qwen-extracted audiovisual fields.

Task	T1	T2	T3	T4	T5
Shared routing cap (chars)	10K	11.5K	13.5K	16K	8.5K

Table 12: Task-matched pre-tokenization character routing limits used by ASR-Retrieval and KnowMe. For T4, both routes use approximately 16K characters.

offline memory corpus indexes 327 source videos as 2,348 clip units and totals 549,680 tokens across clip, segment, persona, and trajectory layers. Per-unit token cost averages 168 tokens for clip evidence, 49 tokens for segment summaries, and 51 tokens for theme nodes.

The controlled quantity is therefore routed characters before answerer tokenization, not exact final-token equality. Because KnowMe routes compressed summaries while ASR-Retrieval routes transcript-derived chunks, comparable character caps yield different token counts. Across T1–T5, total per-question KnowMe input including routed evidence, prompt, and instructions is approximately 1.8K–2.8K tokens, compared with about 8K–12K tokens for ASR-Retrieval and about 28K for ASR-Flat-32K. The shorter KnowMe context reflects compression rather than a larger input allowance.

H Human Performance Reference

The human-reference row in Table 3 uses the same prompts, source-identifier interface, and scoring protocol as the evaluated systems. The same five graduate-student T4 question-authors serve as annotators, one per source series; each answers a series they did not author. Per series, the set contains 20 items each for T1, T2, and T5 and 30 each for T3 and T4, yielding aggregate denominators of 100/100/150/150/100. Human-reference scores are 92.0/88.0/82.0/78.0/93.0 on T1–T5.

I Artifact Availability and Rights Status

The five source collections are public personal-vlog series hosted on YouTube. Raw videos are not redistributed. At the camera-ready stage, we have not obtained blanket creator permission to redistribute benchmark annotations, gold answers, ASR segments, extracted frames or frame-derived fields, generated memories, retrieval outputs, or other source-derived artifacts. We are contacting the original creators to clarify these permissions. We therefore make no blanket public-release commitment for benchmark data or source-derived materials and assign no blanket data license in this paper. Any future availability must be determined artifact-by-artifact and source-by-source under the permissions obtained and applicable platform, privacy, and rights constraints. Author-created evaluation code is separable from the source-content rights issue, but this paper likewise does not make a public-release timing commitment for it.

J Additional Qualitative Diagnostics

Qualitative inspection focuses on two representative failure modes. On T4, transcript retrieval often surfaces one explicit mention and produces a narrow answer, while KnowMe combines repeated evidence across videos into a stable pattern. On T3, failures often arise when summaries preserve the broad topic but remove the local contrast needed for the correct ordering. These cases motivate the separation between ordering and synthesis in the main benchmark taxonomy.