X-LeBench: A Benchmark for Extremely Long Egocentric Video Understanding

Anonymous ACL submission

Abstract

Long-form egocentric video understanding provides rich contextual information and unique insights into long-term human behaviors, holding significant potential for applications in embodied intelligence, long-term activity analysis, and personalized assistive technologies. However, existing benchmark datasets primarily focus on single, short (*e.g.*, minutes to tens of minutes) to moderately long videos, leaving a substantial gap in evaluating extensive, ultra-long egocentric video recordings. To address this, we introduce X-LeBench, a novel benchmark dataset meticulously designed to fill this gap by focusing on tasks requiring a comprehensive understanding of extremely long egocentric video recordings. Our X-LeBench develops a life-logging simulation pipeline that produces realistic, coherent daily plans aligned with real-world video data. This approach enables the flexible integration of synthetic daily plans with real-world footage from Ego4D—a massive-scale egocentric video dataset covers a wide range of daily life scenarios—resulting in 432 simulated video life logs spanning from 23 minutes to 16.4 hours. The evaluations of several baseline systems and multimodal large language models (MLLMs) reveal their poor performance across the board, highlighting the inherent challenges of long-form egocentric video understanding, such as temporal localization and reasoning, context aggregation, and memory retention, and underscoring the need for more advanced models. Our dataset is available at X-LeBench.

1 Introduction

Understanding long-form egocentric videos captured from a first-person perspective over extended periods holds significant potential for advancing various domains such as embodied intelligence, long-term activity analysis, and personalized assistive technologies (Plizzari et al., 2024; Lv et al., 2024; Park et al., 2016). These videos provide rich contextual information and unique insights into human behaviors as they unfold naturally throughout the day. The ability to analyze such extensive recordings is key to developing more personalized agent systems that can construct long-term memory, anticipate user needs, and interact seamlessly in real-world settings (Lin et al., 2022b; Jia et al., 2022; Pramanick et al., 2023).

However, most existing datasets (Soomro, 2012; Caba Heilbron et al., 2015; Kay et al., 2017; Li et al., 2020; Xiao et al., 2021; Miech et al., 2019), predominantly feature short individual video clips captured from third-person views, making them insufficient for understanding the continuous, nuanced context of daily human activities or for in-depth human-centered research that requires a first-person perspective. While more recent benchmarks and datasets focus on egocentric video understanding (Damen et al., 2018, 2022; Grauman et al., 2022; Zhu et al., 2023; Mangalam et al., 2023; Grauman et al., 2024; Lv et al., 2024; Fan et al., 2025), they remain limited in capturing continuous, long-term human activities. For instance, Ego4D (Grauman et al., 2022) features massive-scale egocentric videos covering a wide range of daily activities, with video durations ranging from 5 seconds to 7 hours, yet subsequent research (Mangalam et al., 2023; Grauman et al., 2024; Islam et al., 2024; Rodin et al., 2024; Chandrasegaran et al., 2024) often focuses on isolated clips or recordings, falling short of capturing the full scope of long-term daily human activities. This limitation hinders the evaluation of models designed to process ultra-long video streams. In particular, it restricts the ability to challenge current long-form video understanding systems and models that construct long-term memory from video recordings and retrieve relevant information in response to user queries. Without extensive, continuous contextual data, evaluating the robustness of their performance becomes difficult.

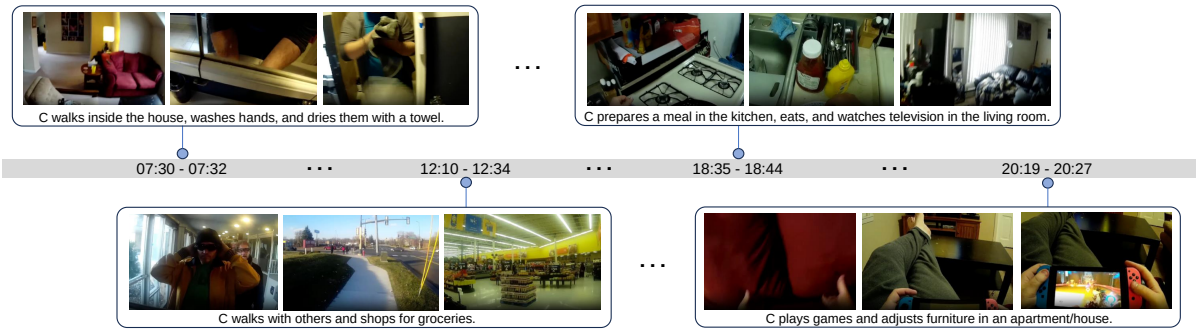

Figure 1: Example of generated video life logs in X-LeBench. Generated video life logs consist of multiple videos with corresponding timestamps. The visualization shows the time organization and content allocation of data.

Creating benchmark datasets that span several hours of egocentric video is necessary but presents significant challenges. Data acquisition is a primary hurdle, requiring participants to wear recording devices for extended periods is labor-intensive and raises privacy concerns. Device limitations, including storage constraints and reliability issues, further complicate continuous video capture. Additionally, annotating long-form videos is time-consuming and prone to annotator fatigue, affecting label accuracy and consistency.

To address these challenges, we introduce X-LeBench, a versatile and scalable benchmark dataset designed for evaluating tasks on extremely long egocentric videos. X-LeBench features a life-logging simulation pipeline that simulates extended video logs by integrating short (seconds) or moderately long (hours) video clips with dynamically generated daily plans. Leveraging large language models (LLMs), it generates realistic, contextually rich schedules aligned with real-world activities based on adjustable input settings. Specifically, this simulation integrates synthetic daily plans with actual footage from Ego4D, then iteratively optimizes the simulation process based on retrieved information, producing video life logs that mirror daily activities in rich contexts with duration extended to dozens of hours. A generated sample is shown in Fig. 1, showcasing video segments with different timestamps. Notably, X-LeBench offers a customizable and scalable design, enabling the synthesis of datasets with various durations and content to accommodate diverse research needs.

Our initial evaluations of baseline systems and multimodal LLM (MLLM) reveal consistently poor performance on X-LeBench, highlighting the inherent difficulties of long-form egocentric video understanding and underscoring the need for more advanced models capable of interpreting and analyzing ultra-long egocentric videos.

Our contributions are: (1) We present X-LeBench, the first benchmark dataset that encompasses ultra-long egocentric video recordings; (2) We introduce a novel and customizable pipeline that simulates realistic, hours-long egocentric video life logs by integrating synthetic daily plans with real-world footage; (3) We conduct extensive evaluations of existing models on X-LeBench, exposing significant performance gaps and key challenges for future research.

2 Related Works

2.1 Egocentric Video Benchmarks

Egocentric video understanding has received increasing attention (Cheng et al., 2024; Huang et al., 2024). Ego4D is a landmark dataset, offering 3,670+ hours of egocentric footage across 74 locations, facilitating significant progress in the field.

Related works (Bain et al., 2023; Lin et al., 2022a; Pramanick et al., 2023; Islam et al., 2024) extend Ego4D, further exploring various applications of egocentric video understanding. EgoSchema (Mangalam et al., 2023) offers an egocentric video question-answering benchmark with over 5,000 curated multiple-choice pairs but focuses on 3-minute clips. The AEA dataset (Lv et al., 2024) offers multimodal egocentric data, HourVideo (Chandrasegaran et al., 2024) curates hour-long videos from Ego4D for evaluating video-language understanding. EgoPlan-Bench2 (Qiu et al., 2024) assesses planning capabilities of MLLMs. However, these works either lack coherent, continuous, long-term daily-life recordings or face data scale and diversity limitations.

Recently, EgoLife's week-long recordings (Yang et al., 2025) offer valuable long-context data, but high costs, operational issues, limited 6 subjects, and an indoor focus restrict its scalability and di-

versity. In contrast, we developed a novel, cost-effective pipeline to synthesize ultra-long, coherent egocentric video life logs from existing datasets, offering a scalable and extensible alternative and significantly broadening the scope and applicability of long-term egocentric video understanding.

Table 1: The comparison of various benchmarks.

Dataset	Avg. Duration (mins)	Annotation Scheme	Egocentric	#QAs	#Data
MVBench	0.27	Auto	No	4,000	3,641
ActivityNet-QA	1.85	Manual	No	8,000	800
EgoSchema	3	Auto&Manual	Yes	5,063	5,063
EgoPlan-Bench2	up to 5	Auto&Manual	Yes	1,321	1,113
MovieChat-1K	9.4	Manual	No	1,950	130
MLVU	12	Auto&Manual	Partial	2,593	757
Video-MME (Short)	1.37	Manual	No	2,700	900
Video-MME (Medium)	9.38				
Video-MME (Long)	39.76				
HourVideo	45.7	Auto&Manual	Yes	12,976	500
InfiniBench	76.34	Auto&Manual	No	108,200	1219
LVBench	68.35	Manual	No	1,594	103
EgoLife	2,658	Auto&Manual	Yes	3,000	6
Ours (Short)	142	Auto&Manual	Yes	26,932	432
Ours (Medium)	319				
Ours (Long)	516				

2.2 Long-form Video Benchmarks

The definition of "long" in video understanding varies across benchmarks, with durations ranging from minutes to hours. As shown in Tab. 1, EgoSchema defines 3-minute videos as long using their proposed certificate length, Video-MME (Fu et al., 2024) classifies videos with a length of 30-60 minutes as long. HourVideo and LVBench (Wang et al., 2024b) define long videos as 20+ and 30+ minutes, respectively. MLVU (Zhou et al., 2024) offers videos of diversified lengths with a 12-minute average length, providing comprehensive evaluation tasks for MLLMs' long video understanding capabilities. InfiniBench (Ataallah et al., 2024) pushes the boundary with 50-minute videos and over 108,000 question-answer pairs, posing significant challenges for leading AI models.

Despite recent advances, existing datasets remain insufficiently long for evaluating ultra-long video processing and rarely focus on egocentric content. X-LeBench fills this gap by redefining "long video" to include multi-hour, contextually consistent egocentric recordings with rich annotations to advance model and system development.

2.3 LLM-assisted Annotation Scheme

Traditional annotation processes require extensive human effort, but advances in LLMs and MLLMs (Achiam et al., 2023; Yao et al., 2024; Bai et al., 2023; Team et al., 2023; Touvron et al., 2023; Lin et al., 2023) have facilitated automated annotation in benchmarks. As shown in Tab. 1, several video understanding benchmarks leverage LLMs/MLLMs to streamline annotation process (Mangalam et al., 2023; Rawal et al., 2024; Pătrăucean et al., 2023; Li et al., 2024). For instance, EgoSchema generates question-answer pairs by querying an LLM with human-annotated narrations and tailored prompts.

Given the multi-hour duration of our dataset, manual annotation is impractical and prone to fatigue-induced errors. To address this, we also adopt an automatic and manual annotation scheme. First, we adapt Ego4D's manual annotations to align with our task requirements. Next, we use LLMs to consolidate summaries of 5-minute video clips from Ego4D at various granularities, creating single-video, multi-video, and holistic-level summaries (detailed in Appendix C).

3 X-LeBench

In real-world scenarios, continuous egocentric life-logging is often constrained by hardware limitations and privacy concerns, making it infeasible to record uninterrupted long-term videos. Consequently, long-duration life-logging must be reconstructed from multiple video segments captured at different times. Building on this insight, we develop the life-logging simulation pipeline that leverages LLMs' text-processing capabilities and Ego4D's extensive annotations. It generates simulated video life logs that realistically reflect daily activities and maintain contextual coherence, leading to the creation of X-LeBench. The following sections detail our methodology.

3.1 Life-logging Simulation Pipeline

As shown in Fig. 2, we constructed the pipeline by implementing the following three stages:

Stage 1 - Persona Generation. To ensure the simulations better reflects the multifaceted nature of real-world human activity, enriching the realism and variability of the resulting dataset, we dynamically generate personalized character profiles based on different predefined locations and the Myers-Briggs Type Indicators (MBTI) (Myers, 1962), generating basic background information including the character's personality traits, lifestyle, hobbies and general daily routines. By incorporating varied character settings, we capture a wide range of behavioral variations that enrich our dataset. This pragmatic choice enables the generation of diverse daily plans and activity patterns, promoting variability across simulations.

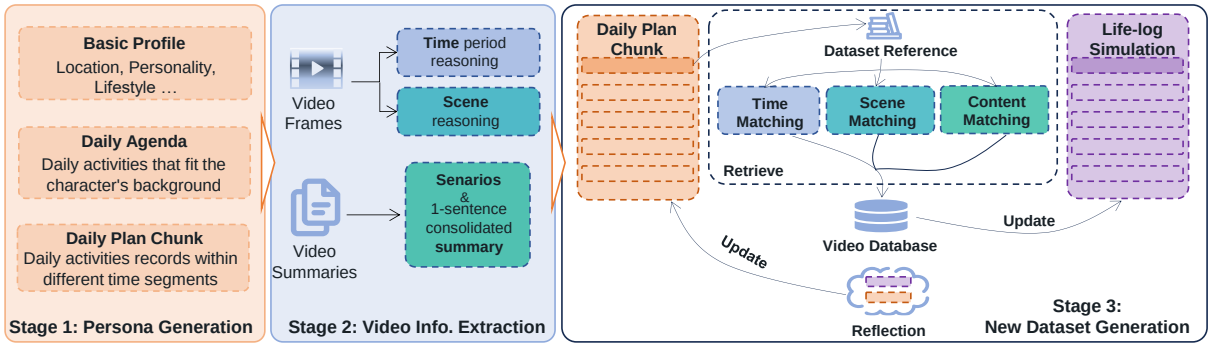

Figure 2: Overview of life-logging simulation pipeline. Stage 1: Generation of personalized persona profiles and daily plan chunks based on predefined parameters. Stage 2: Core information (time, scene, content) extraction of videos in the video selection library. Stage 3: Matching and retrieval of daily plan chunks with videos, and life-log simulations are iteratively refined through reflection, resulting in the final optimized output.

Each daily plan is then segmented into time-specific activity chunks, ensuring structured activity distribution. During this stage, we use GPT-4o (Hurst et al., 2024) for character generation, setting 9 different locations based on Ego4D's demographic data, and 16 MBTI types per location, resulting in 144 diverse persona profiles (profile example is shown in Appendix A).

Stage 2 - Video Information Extraction. We construct our video selection library by carefully selecting 7852 videos from the Ego4D labeled with scenario information and dense clip-level summaries, excluding redacted content. Then we extract three key attributes from each video, including the time, scene, and the main content. Specifically, we use Gemini-1.5-Pro (Team et al., 2024) to analyze videos and transform information. To facilitate simulation with reasonable environment settings and lighting conditions, videos are categorized into different scene environments: indoor, outdoor, and mixed, as well as different time periods: daytime, nighttime, twilight, and uncertain (*e.g.*, indoor videos lacking external cues are labeled "uncertain"). Content information is derived from Ego4D scenario tags, and 1-sentence summaries consolidated from 5-minute clip-level summaries from each video. This structured metadata ensures precise alignment between video content and the generated daily plans in the next stage.

Stage 3 - New Dataset Generation. In this stage, we match the generated daily plan chunks (Stage 1) with videos (Stage 2) to construct coherent video life-log simulations. For each plan chunk, we retrieve a video that aligns with its time, scene, and content, the selection process follows below rule:

$$V = \underset{V \in \mathbf{M_v}}{\arg\max}\, S(\mathbf{M_v}(C_t, C_{scene}, C_{scenario}), C_{desc}) \quad (1)$$

where V is the selected video for chunk C, determined by the alignment with C's core information, which is time (C_t), scene (C_{scene}), scenario ($C_{scenario}$), and 1-sentence summary description (C_{desc}). $\mathbf{M_v}(\cdot)$ denotes the operation of matching videos based on the provided chunk information. $S(\cdot)$ represents the sentence similarity computation between C_{desc} and 1-sentence summary set of matching videos. Specifically, C_t and C_{scene} are inferred from chunk timestamps and content, while $C_{scenario}$ is inferred using the dataset's predefined scenario list (*i.e.*, the dataset reference in Fig. 2).

To efficiently identify most suitable videos from the large-scale video selection library, selection process follows a coarse-to-fine retrieval strategy. We first apply a coarse-grained filtering step, *i.e.*, the operation $\mathbf{M_v}(C_t, C_{scene}, C_{scenario})$, a rule-based matching to narrow down the candidate video set. Finally, to refine the selection, we calculate the sentence similarity between C_{desc} and the summaries of the candidate video set $\mathbf{M_v}$, ensuring alignment across time, scene, and content dimensions.

In addition, to enhance the contextual reasonableness and coherence of overall video life logs, we preferentially select videos recorded in the same location for each simulation. Furthermore, an iterative optimizing strategy is adopted, *i.e.*, after each matching of a daily plan chunk, we update the life-log simulation memory with the information of selected video, require LLM uses this context to reasonably adjust and update the subsequent daily plan chunks, ensuring that the overall simulation maintains logical coherence and contextual alignment. Details can be found in Appendix B.

Statistics. Our life-logging simulation pipeline allows for flexible customization by inputting various location and MBTI type, generating diverse

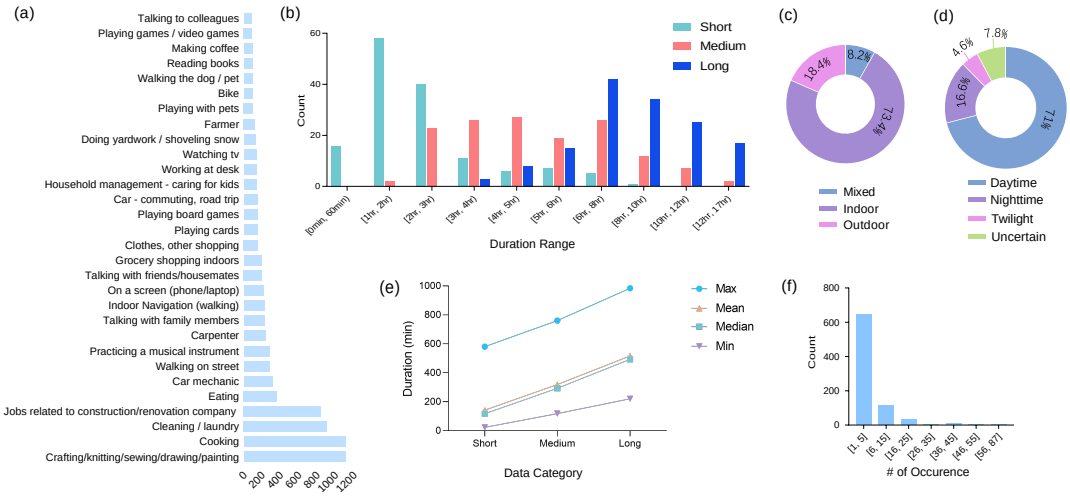

Figure 3: Dataset statistics. (a) Video selection library consists of 7852 videos from Ego4D dataset, covering 135 scenarios, here shows the top 30. (b) X-LeBench has video life-log durations span from 23 minutes to 16.4 hours, here shows distribution of duration range across different data categories. (c) and (d) show the distribution of scene and time information categories of videos in video selection library, respectively. (e) Statistical information of duration lengths of different data categories, including minimum, mean, median and maximum values. (f) The distribution of number of videos with different occurrences.

persona profiles and simulations. We can customize different numbers of chunks to get video life logs of various video compositions. In this work, we set three different chunk numbers for simulation, namely 4, 9, and 15 (corresponding to short, medium, and long life-log categories) for each location-MBTI combination generated persona, thus obtaining 432 extremely long video life logs. To assess dataset quality, we compare our method with a randomly sampled baseline (Appendix E), our method retrieves videos with contextualized timestamps and more consistent content that closely resemble real-life recorded scenes. We also conduct a five-point human evaluation on realism and contextual consistency. As summarized in Tab. 2, Fig. 3 (b) and (e), our dataset achieved scores over 4 across all categories, demonstrating strong alignment with real-world activity patterns and logical continuity. Evaluation details are provided in the Appendix D. Also, the dataset exhibits diverse duration distributions: Short data are mostly 1-2 hours; Medium data are mostly 4-5 hours; And long data are mostly 6-8 hours. With the average duration of 2.37, 5.32 and 8.6 hours.

Table 2: Statistics of dataset duration length and quality.

Life-log Category	Max (min)	Mean (min)	Min (min)	Realism	Contextual Consistency
Short	580	142	23	4.71	4.50
Medium	760	319	118	4.26	4.37
Long	984	516	220	4.41	4.02

Our video selection library contains 7852 videos, covers a wide range of daily life scenarios and scenes across different times of the day, a total of 135 scenarios are covered. Fig. 3 (a), (c) and (d) illustrate the top 30 most frequent scenarios, time distributions, and scene types. In addition, due to the commonality of human activities (*e.g.*, eating, cooking), certain videos are selected multiple times. Fig. 3 (f) shows that most videos appear fewer than five times, with a small number of videos appearing more than 5 times, reflecting the diversity of our simulations while preserving the representativeness of repeated contexts.

Table 3: The number of annotations.

Object-related Retrieval	People-related Retrieval	Action Counting	Summary Ordering
1444	583	5295	4032

Moment Retrieval	Summarization		
	single-video	multi-video	holistic
9869	4032	1245	432

3.2 Benchmark Tasks

As shown in Tab. 4, X-LeBench introduces a suite of evaluation tasks specifically designed for daily-life long-form egocentric videos. These tasks encompass object-, people-, and moment-related temporal localization, multi-level summarization, action counting, and summary ordering. Details and examples are in Appendix C.

The full-length video is presented to the system only once before querying. The system is required to extract features or frames from the extremely long video and store them in a buffer, responding to query based solely on the stored information.

Temporal Localization		
Object-related Retrieval	Q: What did I put on the table in the record provided for the 21:56 - 22:25 time period?	A: [22:01:55 - 22:02:03], [22:02:20 - 22:02:42], ...
People-related Retrieval	Q: Who did I talk to in the living room in the record provided for the 21:56 - 22:25 time period?	A: [22:02:35 - 22:02:39], [22:03:30 - 22:03:50], ...
Moment Retrieval	Q: When did I ⟨use phone⟩ in the record provided for the 21:56 - 22:25 time period?	A: [22:15:00 - 22:15:06], [22:18:49 - 22:19:00], ...
Summarization		
Single-video Summarization	Q: Summarize the activities performed in the recordings provided for the time period 22:13 - 22:25.	A: C interacts with others and writes in a house and a studying room.
Multi-video Summarization	Q: Summarize the activities performed in the recordings provided in the morning (before 12:00).	A: C engages in various activities, including preparing meals, riding...
Holistic Summarization	Q: Summarize the activities performed in all provided life-log recordings.	A: C engages in a variety of activities throughout the day, including preparing meals, riding, visiting restaurants, ...
Counting		
Action Counting	Q: In the record provided for the 07:25 - 07:48 time period, how many times have each of the actions in the following list been performed in the 0-28 second record of the period? Action list: 1. wipe soap, 2. remove cheese, ...	A: wipe soap: 2; remove cheese: 1; ...
Ordering		
Summary ordering	Q: Please rank the following summaries of camera wearer C's activities in order of presentation of the life-log recordings. Summary 0: C climbs, interacts with others, and observes climbing activities in various indoor... ... Summary 14: C interacts with others and writes in a house and a studying room.	A: Correct order of the summaries: 7, 2, 11, 5, 8, 1, 6, 3, 4, 0, 12, 9, 13, 10, 14.

Table 4: Tasks in this benchmark and their corresponding examples. **Q** and **A** denote query and answer examples.

This approach presents a significant challenge to current video understanding systems, which are typically designed to process short videos and thus cannot store information in the presented videos in buffers for future use. Furthermore, as video duration increases, answers to queries may recur across multiple time segments. To address this, we explicitly annotate the queried period for each task, ensuring precise and unambiguous retrieval, while mitigating annotation gaps inherent in single-video datasets. After generating the life-log footage, we adapt corresponding Ego4D annotations for our novel task designs. Tab. 3 summarizes the number of annotations per task.

4 Experiments

4.1 Settings

X-LeBench includes various task types, as outlined in Sec. 3.2, encompassing four categories: temporal localization, summarization, counting, and ordering. These tasks are further divided into eight subtasks, with their respective input-output formats detailed in Tab. 4. Given the interleaved, ultra-long chronological context (Fig. 1), we uniformly construct time-stamped prompts as the long-context input following the format shown in Fig. 5.

To reflect real-world usability, tasks require free-form textual outputs instead of closed-set multiple-choice answers. This approach increases practical usability while significantly raising the complexity of tasks, as it demands more nuanced understanding and reasoning. To ensure the evaluability of outputs, the prompts provide extended context along with task-specific instructions and output format requirements for each task type, standardizing the expected responses. For temporal localization and summarization, each query is independently evaluated, while for counting and ordering, we aggregate the query contents of each data entry to action or order lists for a unified assessment. Notably, the computational cost and time associated with analyzing ultra-long contexts, along with their overall poor performance and limited analytical value, we test only a limited set on temporal localization.

Overall, X-LeBench presents significant challenges for current video understanding systems, which are often limited by short context handling, weak temporal reasoning, and poor long-horizon memory retrieval. To evaluate performance under these constraints, we test three representative multimodal approaches to handle ultra-long contexts:

(1) Gemini-1.5-Flash (Team et al., 2024): A native multimodal model excels in handling ultra-long contexts and is trained jointly on multimodal data. It is used in an end-to-end manner by uniformly sampling 1200 frames per data, with a temperature of 0.1.

(2) Socratic Models (Zeng et al., 2022): Most advanced multimodal systems cannot process ultra-long videos. Inspired by related works (Chandrasegaran et al., 2024; Zeng et al., 2022), we divide videos into 30-second segments and generate captions for these segments using the Qwen-VL-7B model (Wang et al., 2024a), captioning segments at 1 frame per second and resolution settings of 560×420. The captions are then aggregated with timestamps as life-log records, forming the input for executing long-context understanding tasks. The Gemini-1.5-Flash model is used for performing tasks on these textual life log records, with the same parameters as in method (1).

(3) Retrieve-Socratic: Considering the inherent temporal nature of our dataset, we propose an enhanced version of the Socratic Models approach. This method incorporates a rule-based filtering mechanism to extract only task-relevant

time segments, reducing irrelevant temporal data. By narrowing the contextual focus, this approach aims to improve efficiency and relevance in long-context understanding. The extracted contextual information is then processed using the same Socratic Models' pipeline and parameter settings as in method (2).

We also evaluate the LongVU (Shen et al., 2024), an open-source method for long-form video understanding, detailed results are in Appendix E.

4.2 Results

As shown in Fig. 4 and Tab. 5, we evaluate the performance of three methods across short, medium, long, and all data categories. The "all" category represents an aggregated evaluation of the former 3 data types, offering a more comprehensive and holistic assessment. The evaluation metrics for each task are detailed in Appendix C.

Overall Performance. As shown in last three rows of Tab. 5 and Fig. 4 (a), the Retrieve-Socratic method demonstrates superior performance in most tasks. Due to input token limitations, longer videos require larger frame sampling intervals, resulting in increased information loss. Consequently, both Socratic and Retrieve-Socratic methods significantly outperform Gemini-1.5-Flash on temporal localization and summarization tasks. Specifically, Retrieve-Socratic achieves an average improvement of 8.26% in recall for temporal localization and 1.87 points higher scores in summarization. The underperformance of Gemini-1.5-Flash in temporal localization reflects its difficulty in fine-grained temporal reasoning under long-context constraints. Additionally, Retrieve-Socratic further improves upon the basic Socratic approach by narrowing the temporal context, leading to notable gains in both temporal localization and summarization tasks. In holistic summarization and ordering, where the queried information spans the entire video, the performance of Socratic and Retrieve-Socratic methods remains comparable, demonstrating their robustness in global reasoning.

Interestingly, our analysis of the specific results indicates that the Retrieve-Socratic method tends to output 0 more frequently in the counting task. This leads to noticeably poorer performance compared to other methods, suggesting that excessive context reduction may introduce unintended biases in certain tasks.

Impact of Data Types. As illustrated in Fig. 4 (b) and Tab. 5, the performance trends across data types are not uniform. These trends may be influenced by varying task-specific objectives and understanding burden imposed by different data types on each method. Overall, both Socratic and Retrieve-Socratic demonstrate more stable performance trends across data types compared to Gemini-1.5-Flash. This suggests that the processing capability of textual information is more robust when dealing with such long-context information. For the ordering task, all methods achieve strong performance on short videos (accuracy exceeding 85%). However, as video length increases, performance across all methods declines significantly, with accuracy dropping below 25% for long data.

Remarks. (1) Model refusal rate: We observe a notable refusal rate from Gemini-1.5-Flash due to its built-in information security constraints, particularly when directly processing long-form video input. The refusal rates increase with input length, reaching 20.14%, 27.08%, and 30.56% for short, medium, and long videos, respectively. both the Socratic and Retrieve-Socratic methods, which convert video content into textual descriptions before prompting the model, achieve a 0% refusal rate. LongVU does not exhibit any refusal behavior. (2) Temporal reasoning failure: In the multi-video summarization task, a significant portion of responses are invalid, even when timestamps are included in input context, and the query time is explicitly specified in the question (*e.g.*, "12:00 to 17:00"). Typical invalid responses include statements like "There is no information available within the specified time period", indicating a breakdown in the temporal understanding of the model. The failure rates of the multi-video summarization task are 55.28%, 51.89%, and 57.03% for Gemini-1.5-Flash, Socratic Models and Retrieve-Socratic, respectively. LongVU demonstrates almost no temporal awareness, it consistently disregards the specified time window and instead summarizes the entire video content. Also, in the ordering task, its performance on short, medium, and long data drops drastically to 37.5%, 1.7%, and 0.69%, respectively. Manual review of the outputs reveals that even on short data, LongVU frequently returns meaningless outputs like "0, 1, 2, 3" without establishing any meaningful temporal structure. Therefore, we only included LongVU in tasks that require global summaries or coarse-grained time reasoning (*i.e.*, holistic summarization and ordering). Additional results and analysis are in the Appendix E.

Method	Data Category	Object-related Retrieval (%)	People-related Retrieval (%)	Moment Retrieval (%)	Summarization (10)			AVG. Temporal Localization (%)	AVG. Summarization (10)	Action Counting (%)	Ordering (%)
					single	multi-video	holistic				
Gemini-1.5-Flash	Short	1.67	0.00	10.26	5.12	3.02	4.90	4.20	4.36	13.87	93.70
	Medium	1.92	6.90	10.53	4.82	3.12	5.24	5.88	4.46	16.74	44.66
	Long	1.75	11.11	8.33	4.57	3.44	5.33	6.45	4.43	16.76	24.40
Socratic Models	Short	8.33	25.00	17.95	6.80	3.69	6.47	14.28	5.67	9.53	85.59
	Medium	5.77	20.69	18.42	6.64	4.09	6.41	13.45	6.04	14.04	55.63
	Long	12.28	22.22	10.83	5.82	4.37	6.15	11.83	5.61	19.85	23.33
Retrieve - Socratic	Short	8.33	25.00	17.95	6.88	3.79	6.53	14.28	5.75	9.92	87.67
	Medium	9.62	17.24	7.89	7.01	4.30	6.56	10.92	6.36	9.38	54.55
	Long	10.53	66.67	14.17	6.84	4.69	6.48	15.59	6.48	9.77	24.17
Gemini-1.5-Flash	All	1.77	5.17	9.14	4.73	3.20	5.16	5.66	4.43	16.44	41.96
Socratic Models	All	8.87	22.41	13.70	6.22	4.06	6.34	12.97	5.76	**17.15**	42.61
Retrieve - Socratic	All	**9.47**	**27.59**	**13.71**	**6.90**	**4.28**	**6.52**	**13.92**	**6.30**	9.67	**43.01**

Table 5: Evaluation results on X-LeBench. Including temporal localization tasks (object-related, people-related and moment retrieval), summarization tasks (single-video, multi-video and holistic level), action counting and summary ordering task. AVG. temporal localization and summarization is the average performance of all temporal localization tasks and summarization tasks, respectively.

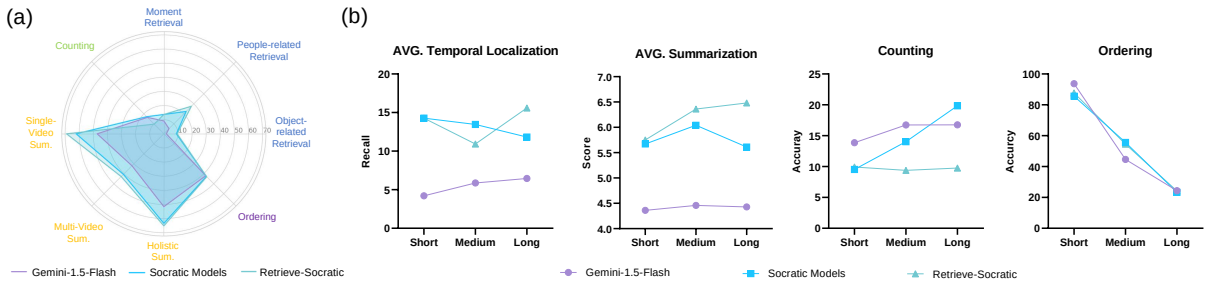

Figure 4: Performance comparison on X-LeBench. (a): Radar chart showing the overall performance of each method across all tasks. For fair comparison across tasks, scores for summarization are scaled by a factor of 10 (maximum 100). (b): The performance comparison across different data categories for different tasks.

4.3 Summary of Key Findings

Temporal Reasoning: The Core Bottleneck. Evaluated models struggle with temporal reasoning, showing high failure rates in multi-video summarization and localization. This highlights the need for models with improved temporal alignment and contextual memory for long sequences.

Textual Representations: A Path Towards Scalable Understanding. Our findings suggest that structured textual representations serve as a highly effective and scalable intermediate for ultra-long video understanding. As demonstrated by the Socratic and Retrieve-Socratic methods, converting raw long video inputs into textual forms significantly reduces model refusal rates and consistently enables more stable and improved performance. This strategy not only effectively alleviates the severe token limitations of current multimodal models but also dramatically reduces computational overhead. This makes it a practical and highly scalable alternative for long-form video understanding, especially in scenarios where direct processing of extended video sequences remains challenging.

Context Filtering: Balancing Efficiency and Completeness. The Retrieve-Socratic approach leverages rule-based temporal filtering to retain only task-relevant context, resulting in superior performance in temporal localization and summarization tasks. However, its relatively weaker performance in counting task suggests that aggressive pruning may omit essential contextual cues. Going forward, this insight underscores the need to design more adaptive retrieval-augmented generation (RAG) systems capable of balancing context filtering with comprehensive long-term information retention in ultra-long video understanding.

5 Conclusion

We introduce X-LeBench, the first benchmark for ultra-long egocentric video understanding, featuring a customizable life-logging simulation pipeline. By curating 432 video life logs from Ego4D, categorized into short, medium, and long durations. X-LeBench provides a diverse, structured dataset for long-form video analysis. It includes tasks in temporal localization, summarization, counting, and ordering, offering a rigorous evaluation framework. We hope to to advance research in long-form egocentric video processing, fostering the development of more robust and temporally aware AI models.

6 Limitations

Despite our efforts to construct an extremely long video dataset with contextually coherent activity contents, limitations remain due to the scarcity and insufficient diversity of available data. For details of videos, contextual inconsistencies exist within the dataset, such as different kitchen or office settings within the same data. However, since our focus is on long-term activity content, these discrepancies are beyond the scope of this work. In addition, while MBTI is employed in our simulation pipeline to introduce behavioral diversity, we acknowledge that it is a heuristic framework with limited scientific validity. Its inclusion serves a pragmatic purpose—to support structured variation in persona-driven daily plans—and is not intended to imply psychological rigor or generalizable personality modeling.

Future work could involve expanding both the modalities and time spans of collected data through diverse devices and sources to enhance a comprehensive understanding of long-form human activity.

References

Hervé Abdi. 2007. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2024. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding. *arXiv preprint arXiv:2406.19875*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23.

Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. 2025. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *European Conference on Computer Vision*, pages 428–448. Springer.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. 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*, pages 19383–19400.

Yifei Huang, Jilan Xu, Baoqi Pei, Yuping He, Guo Chen, Lijin Yang, Xinyuan Chen, Yaohui Wang, Zheng Nie, Jinyao Liu, et al. 2024. Vinci: A real-time embodied smart assistant based on egocentric vision-language model. *arXiv preprint arXiv:2412.21080*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *EMNLP*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022a. Egocentric video-language pre-training. *arXiv preprint arXiv:2206.01670*.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022b. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586.

Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. 2024. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*.

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

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

IB Myers. 1962. The myers-briggs type indicator. *Educational Testing Service/Princeton*.

Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. 2016. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705.

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. 2024. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, pages 1–57.

Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*.

Lu Qiu, Yuying Ge, Yi Chen, Yixiao Ge, Ying Shan, and Xihui Liu. 2024. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, and Giovanni Maria Farinella. 2024. Action scene graphs for long-form understanding of egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18622–18632.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

K Soomro. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. 2025. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv*.

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence*, 44(8):4252–4266.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Chenchen Zhu, Fanyi Xiao, Andrés Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. 2023. Egoobjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20110–20120.

A Persona Profile

To simulate diverse egocentric behaviors, we generated 144 unique persona profiles by combining 16 MBTI personality types with 9 different location settings. Each profile serves as input for our dataset creation. Tab. 6 presents a detailed example of a persona profile, illustrating its structure and content.

B Coarse-to-Fine Video Matching and Iterative Optimization

As described in Sec. 3.1, the life-log simulation pipeline generates multiple daily activity chunks, which are then matched with videos from the video selection library to construct the dataset. Stage 3 follows a coarse-to-fine matching approach based on Equation 1 with the implementation details outlined below:

B.1 Coarse Matching ($M_v(\cdot)$)

For each activity chunk C, key information is extracted, including C_t (time period), C_{scene} (scene type), and $C_{scenario}$ (scenario). The C_{scene} and $C_{scenario}$ attributes are inferred using GPT-4o based on the chunk's textual description C_{desc}. Candidate videos are retrieved from the video selection library where the resource region matches

Field	Description
Location	UK
Personality Traits	• MBTI Type: INFP • Character Traits: – Idealistic and creative, often daydreaming about a better world – Highly empathetic and compassionate, with a strong sense of personal values – Prefers deep, meaningful connections over a wide social circle – Flexible and adaptable, but can be disorganized and easily overwhelmed by details – Values authenticity and is highly attuned to the feelings of others
Lifestyle	C wakes up around 7:30 AM and starts the day with meditation or journaling. Works from home as a freelance writer from 9:00 AM to 5:00 PM, with breaks for lunch and short walks. Evenings are spent reading, writing poetry, or engaging in creative projects, and bedtime is around 11:00 PM.
Hobbies	• Writing poetry and short stories • Reading literature and philosophy • Practicing mindfulness and meditation • Volunteering at local community centers • Exploring nature and taking long walks in the countryside
Daily Agenda	• Wake up at 7:30 AM • Meditation or journaling from 8:00 AM to 8:30 AM • Start work at 9:00 AM • Lunch break from 1:00 PM to 2:00 PM • End work at 5:00 PM • Evening walk or light exercise from 6:00 PM to 7:00 PM • Dinner and relaxation from 7:00 PM to 9:00 PM • Reading or creative activities from 9:00 PM to 10:30 PM • Go to bed at 11:00 PM
Daily Plan Chunks	• 07:30–08:00: Wake up and stretch • 08:05–08:35: Meditation or journaling • 08:40–09:10: Prepare a healthy breakfast and make tea • 09:15–09:45: Eat breakfast, clean the kitchen, and set up the workspace • 09:50–10:50: Start work as a freelance writer • 10:55–11:00: Continue working on writing projects • 11:05–11:20: Take a short break to stretch and hydrate • 11:25–13:00: Continue working on writing projects • 13:05–14:00: Lunch break and light reading • 14:05–16:00: Resume work and handle emails and client communications • 16:05–17:00: Continue handling emails and client communications • 17:05–17:35: End work and tidy up the workspace • 17:40–18:40: Evening walk or light exercise in the park • 18:45–20:00: Prepare and eat dinner, relax with some music • 20:05–21:35: Engage in creative activities (writing poetry, reading literature, or working on a personal project) • 21:40–22:30: Wind down with light reading or calming music

Table 6: Persona Profile: Tabular representation of an INFP individual's daily life, lifestyle, hobbies, and activities.

the persona's predefined location. A video is included in the candidate set $M_v(\cdot)$ when its time and scene type match the C_t and C_{scene}, with an overlap between its scenarios and $C_{scenario}$ larger than 0.33 (*i.e.*, the IoU of their scenarios). Note that, if no matching videos are found, the criteria are relaxed by lowering the overlap threshold to 0.2 or removing the regional constraint. This process produces the initial candidate set $M_v(\cdot)$.

B.2 Fine Matching ($S(\cdot)$)

From the candidate set $M_v(\cdot)$, a finer selection is made by computing the sentence similarity $S(\cdot)$ between the chunk's textual description C_{desc} and the textual summaries of the candidate videos. The two videos with the highest similarity scores are identified, and one is randomly selected to increase dataset diversity while maintaining relevance.

B.3 Iterative Optimization

Once a matching video is selected, its metadata is incorporated into the persona's memory as a record of completed activities. This updated memory, along with the persona's predefined attributes, is provided to GPT-4o to perform a reflection process, enabling the model to refine and adjust the upcoming plan chunks to maintain logical coherence. The next plan chunk is then processed using the updated persona memory, and coarse-to-fine matching is repeated iteratively. This process continues until all plan chunks are matched with corresponding videos, ensuring contextually coherent and realistic life-log simulations.

C Task Details

As described in Sec. 3.2, X-LeBench provides a comprehensive benchmark for daily activity-related tasks, comprising four major categories with eight subtasks. Here, we detail the task settings and the prompt templates used during evaluation. The prompt templates and example outputs are illustrated in Tab. 4 and Fig. 5.

C.1 Temporal Localization

Task Design and Annotation: This task evaluates a model's ability to locate relevant time segments in ultra-long egocentric videos based on user queries. This task, akin to Episodic Memory (Grauman et al., 2022) and Natural Language Video Localization (Zhang et al., 2021, 2020), is uniquely challenging for ultra-long videos due to their extensive duration and diverse content. We divide

Instruction:

You will act as a "Life Recording Assistant" using the provided records to help camera wearer *C* obtain specific information. The provided data contains daily life recordings of C with the corresponding record time, following the format: …

Task Description

First, you need to analyze the given question and identify corresponding segments (i.e., records within the specified or relevant time periods) that contain relevant information or specified in the question. Then, based on your analysis and question, analyze the provided recordings in depth, and get accurate answers to the questions.

Context

The following frames/records are recorded during time of *[start time 1 – end time 1]*:
<records_1>
The following frames/records are recorded during time of *[start time 2 – end time 2]*:
<records_2>
…..

Task-specific Instruction

…..

Figure 5: Example of input context.

this task into three subtasks based on the type of information being retrieved: objects, people, and moments. For object- and people-related retrieval, we adopt the annotations from the Episodic Memory task (specifically, the Natural Language Query subtask) in Ego4D. Specifically, we retrieve the annotations corresponding to the videos selected by the daily plan chunks (as generated in Sec. 3.1), classify them into topics based on the query templates defined in Ego4D, and remapped the timestamps to align with the virtual time used in the daily plan.

While Ego4D defines three query categories—objects, people, and place—we argue that the "Place" query template ("Where did I put X?") fundamentally pertains to object location. Thus, we reclassify these queries under the "object-related" category. Additionally, queries without explicit template annotations were assigned to one of 13 predefined types using an edit-distance-based classification, followed by manual verification. For moment retrieval, annotations for this subtask were derived from the Moments Query subtask of the Episodic Memory task in Ego4D.

Evaluation Metrics: Following previous works on temporal localization (Grauman et al., 2022; Zhang et al., 2021, 2020), we employed a top-x recall with an intersection over union threshold ("recall@x, tIoU"). x and t are predefined variables that indicate the number of retrieved results we look at and the IoU threshold, respectively. Here, we set x to 5 and t to 0.3.

C.2 Summarization

Task Design and Annotation: Summarization is a fundamental task in numerous benchmarks (Wang et al., 2024b; Zhou et al., 2024; Ataallah et al., 2024). In our context, it provides insights at vary-

ing granularities—specific chunks, predefined periods (*e.g.*, morning, afternoon, evening), and holistic summaries. These diverse scopes challenge systems to summarize at multiple summary levels in ultra-long videos. Our summarization tasks involve generating hierarchical summaries for: Single-video (chunk-level) summaries; Multi-video summaries for predefined periods (morning, afternoon, evening[1]); And holistic (full-day) summaries. To generate summaries at these levels, we employed a hierarchical approach inspired by (Islam et al., 2024). Single-video summaries are initially consolidated during the earlier stages of processing (as described in Sec. 3.1). Then, multi-video summaries are produced by LLM-based aggregation, followed by holistic summaries generated in a similar manner.

Evaluation Metrics: We incorporate LLM-based evaluation metrics inspired by MLVU (Zhou et al., 2024) to evaluate our summary tasks from the perspectives of 'Completeness' and 'Reliability'.

C.3 Counting

Task Design and Annotation: This task assesses a model's ability to identify, track, and count occurrences of specific actions across ultra-long videos. The goal is to test fine-grained aggregation capabilities over multiple time segments. Future extensions may include object/person counting.

To maintain annotation reliability, we limit the counting task to intervals with available annotations, as Ego4D annotation only covers a fraction of the total video duration. Unannotated actions could otherwise lead to incorrect counts. Specifically, we derive annotations from the Forecasting task in Ego4D, ensuring consistency.

Evaluation Metrics: We employed a simple accuracy metric. Accuracy was calculated as the ratio of correctly counted actions to the total number of actions queried.

C.4 Ordering

Task Design and Annotation: Ordering task evaluates a model's capability to recognize temporal relationships between main activities across video records in ultra-long video life logs. Unlike prior studies focusing on fine-grained actions, we introduce the novel challenge of ordering events based on single-video summaries. Models are tasked with ordering shuffled single-video summaries of entire life logs, posing significant challenges to the system's temporal understanding ability in long-context inputs. This setup ensures the initial sequence of events is obscured, requiring the model to reconstruct the correct temporal order.

We formulate the task as an open-ended sequence generation problem. Each video summary receives an index post-randomization, preventing numerical hints about the original order. The model is expected to output the correct order based on its understanding of the contents.

Evaluation Metrics: We use ordering accuracy, defined as the ratio between the number of correctly predicted order and the total number of summaries.

D Dataset Quality Evaluation

X-LeBench will be made publicly available, users can access and download videos for research under the Ego4D license (see Ego4D website). To validate the quality of our dataset, we conducted a human evaluation focusing on **realism** and **contextual consistency**. These key aspects are assessed to ensure the dataset accurately reflects realistic human behavior patterns and maintains logical continuity across activities throughout the day. Evaluators concentrated on the content of the activities, disregarding minor inconsistencies (*e.g.*, clothing or environmental details) that do not affect the overall activity realism. The implementation details of this evaluation are outlined in the following sections.

D.1 Evaluation Criteria

The evaluation criteria are defined as follows:

- **Realism.** Each record is assessed for temporal and content alignment with real-world daily activities. Mismatches, such as daytime footage with nighttime timestamps, are considered unrealistic. Evaluators should score all records from 1 (very unrealistic) to 5 (highly realistic).
- **Contextual Consistency.** This criterion evaluates the logical flow and smoothness between consecutive records. Abrupt transitions, such as jumping from office work to outdoor activities, are considered a lack of consistency. Evaluators should score all records from 1 (very poor consistency) to 5 (strong sequence). This ensures the dataset reflects realistic daily human behavior.

[1]Morning: before 12 p.m., afternoon: 12 p.m. to 5 p.m., evening: after 5 p.m.

Table 7: Comparison between our method and random sampling. Left two columns are from our method; Right two columns are from random baseline.

Video Info.	Video Summary	Timestamp	Video Summary	Video Info.
Daytime & Indoor	C washes hands and holds the cupboard door in the bathroom.	06:35	C talks, reads a book, arranges books, draws, and walks inside a house.	Daytime & Indoor
Daytime & Indoor	C, X, Y, and O discuss in a meeting room.	08:35	C scrolls through the phone, climbs a ladder, and fixes cables.	Daytime & Outdoor
Daytime & Indoor	C dials a phone, drinks water, and writes in a bedroom.	16:35	C plays a drum set in a studio.	Nighttime & Indoor
Nighttime & Indoor	C reads a book on the bed in the bedroom.	23:00	C prepares avocado in the kitchen.	Daytime & Outdoor

D.2 Evaluation Setting

Given the extremely long duration of the dataset, which makes full evaluation cost-prohibitive, we select 6 data samples (two each from the long, medium, and short categories), consisting of 56 videos, totalling 31.9 hours. The evaluators are provided with a detailed evaluation manual and the original video data. Each evaluator independently assesses the video records and records their results on a formatted debriefing sheet. Kendall's W (Abdi, 2007) is computed to assess the inter-rater reliability (IRR) of the evaluations.

D.3 Evaluation Procedure

Evaluator Training. Evaluators are trained using detailed guideline (Fig. 6) that include evaluation standards, examples, and formatting instructions.

Evaluation. Three independent evaluators assess all provided data following the guidance. Evaluators score all records from 1 to 5 for realism and contextual consistency, with a brief justification for records scored below 3.

Consensus Evaluation. Compute inter-rater reliability (IRR) using Kendall's W for the scores of 56 records to ensure reliability between evaluators. Resolve discrepancies through discussion and re-evaluation. Here, the evaluations achieve 0.536, it suggests a moderate level of agreement among the raters.

Reporting. The final report presents the mean score for all evaluators.

E Additional Results

E.1 Comparison with Trivial Baseline

Our proposed pipeline is designed to maintain temporal coherence and contextual consistency by integrating multi-dimensional information (time, scene, and content) via life-logging simulation pipeline. To validate the effectiveness of our method, we compare it with a trivial baseline that randomly samples videos. Table 7 shows the comparison between the two methods.

As can be observed, our approach retrieves videos with contextually appropriate timestamps and consistent content, closely mirroring real-life life-logging scenarios. In contrast, random sampling results in time-inconsistent and semantically disjointed videos, further validating the effectiveness of our method.

E.2 Evaluation on LongVU

To extend our evaluation, we tested LongVU (Shen et al., 2024), a state-of-the-art open-source model tailored for long video understanding. Results are summarized in Table 8.

Table 8: Performance of LongVU.

Data Category	Holistic Summary (10)	Ordering (%)
Short	5.21	37.5
Medium	4.81	1.70
Long	4.73	0.69

Our manual review of outputs shows LongVU struggles with temporal reasoning. Even with explicit cues (*e.g.*, "frames 0–500 are recorded from 13:00–13:10"), it still outputs summaries covering the full video. In ordering tasks, LongVU frequently outputs invalid sequences such as "0, 1, 2, 3" or "100, 100, 100, 100", especially for long videos. Consequently, we only evaluated LongVU on tasks that require global information or coarse-grained temporal understanding (*i.e.*, holistic summarization and ordering). Its performance significantly drops as video length increases, underscoring the limitations of current open-source models in ultra-long video understanding and temporal reasoning.

Guideline

1. Your goal is to assess the realism and contextual consistency of given video life-log records, assess if the data accurately reflects realistic human behavior patterns and maintains logical continuity across activities throughout the day.
2. Focusing only on **content-related activity** without considering detailed aspects such as *clothing, gender or house layout consistency.*

1. Evaluation Materials

1. **Dataset**: 6 set video records (2 long, 2 medium, 2 short; total xx hours).
2. **Guidelines**: This manual includes examples and explanations for the evaluation process.
3. **Debriefing Sheet**: Formatted for 5-score rating and justifications.

2. Criteria Definition

Realism: Each record should be assessed for *temporal and content alignment* with real-world daily activities. Mismatches, such as daytime footage with nighttime timestamps, are considered unrealistic. Please score all records from 1 (very unrealistic) to 5 (highly realistic).

- 1 point: Extremely unrealistic in timing or content.
- 2 points: Major inconsistencies with typical human habits or time setting.
- 3 points: Moderately inconsistent, yet somewhat realistic.
- 4 points: Mostly aligns with human routines and timing.
- 5 points: Fully realistic.

Contextual Consistency: This criterion evaluates the *logical flow and smoothness between consecutive records*. Abrupt transitions, such as jumping from office work to outdoor activities, are considered a lack of consistency. Please score all records from 1 (very poor consistency) to 5 (strong sequence).

- 1 point: Poor logical flow with extreme inconsistencies.
- 2 points: Significant unrealistic shifts in activity.
- 3 points: Moderate inconsistencies in sequence.
- 4 points: Mostly logical with minor discrepancies.
- 5 points: Fully coherent and contextually seamless.

3. Evaluation Procedure

Step 1: **Preparation**

- Review this manual thoroughly.
- Familiarize yourself with evaluation standards and examples.

Step 2: **Evaluation**

- **See the criteria definition section first.**
- **Realism Assessment**: For records in each data, rate from 1 to 5, provide a brief justification for records scored below 3.
- **Contextual Consistency Assessment**: For records in each data, rate from 1 to 5, provide a brief justification for records scored below 3.

Figure 6: The content of guideline.