LongEgoRefer: A Benchmark for Long-Form Egocentric Video Referring Expression Comprehension

Shunya Kato[1⋆], Taiki Miyanishi[2], Shuhei Kurita[3,4,5], Mahiro Ukai[4], Nakamasa Inoue[4], and Chenhui Chu[5,6]

[1] Woven by Toyota, Japan
[2] The University of Tokyo, Japan
[3] National Institute of Informatics, Japan
[4] Institute of Science Tokyo, Japan
[5] NII LLMC, Japan
[6] Kyoto University, Japan

Abstract. Egocentric videos capture rich and diverse human–object interactions and have emerged as a fundamental resource for understanding human activities related to objects. In this context, Video Referring Expression Comprehension (Video REC), the task of localizing the temporal and spatial extent of a referred object in video frames given a natural language query, plays a key role in linking textual descriptions to observed objects in untrimmed egocentric recordings. However, existing egocentric Video REC benchmarks primarily focus on short video clips, where some target object appears densely within frames. Such settings do not reflect real-world egocentric recordings, which are long-form, untrimmed, and characterized by sparse object occurrences and complex activity transitions. To address this limitation, we introduce LongEgoRefer, a novel and challenging benchmark constructed from long-form videos in the Ego4D dataset. LongEgoRefer contains 1,498 referring expressions with an average video duration of 45 minutes. The benchmark exhibits extreme target sparsity, detailed linguistic descriptions, and complex human–object interactions embedded in long, dynamic egocentric narratives. Consequently, it defines a demanding spatio-temporal grounding problem that requires models to identify both when an event occurs and where the referred object appears within extended video sequences. We evaluate existing Video REC approaches, including training-free baselines based on vision–language models combined with Grounded SAM2. Extensive experiments show that even advanced baselines and current state-of-the-art models struggle significantly on LongEgoRefer. These results highlight the intrinsic difficulty of long-form egocentric spatio-temporal grounding and emphasize the need for more robust video understanding models. Our benchmark and code are available at `https://github.com/shunya-kato/LongEgoRefer`.

⋆ Woven by Toyota was not involved in this work. The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the official policy or position of Woven by Toyota.

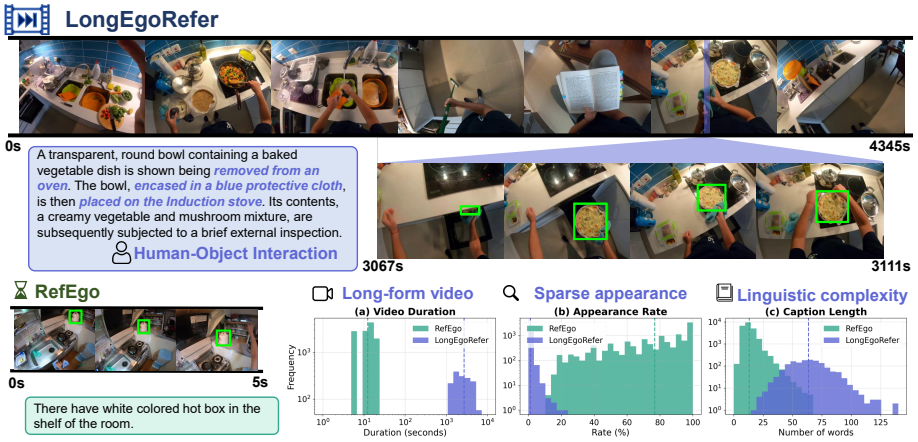

Fig. 1: Data comparison of LongEgoRefer and RefEgo. (a) Video durations in LongEgoRefer are orders of magnitude longer than those in RefEgo. (b) The appearance rate is significantly lower and sparser in LongEgoRefer compared to RefEgo. (c) To handle the complexity of long-form videos, captions in LongEgoRefer are substantially longer and more descriptive than those in RefEgo. Dashed lines indicate the mean value for each distribution.

1 Introduction

Video Referring Expression Comprehension (Video REC) is a critical task in computer vision that involves localizing a specific object or region within a video based on a natural language query, known as *referring expression*. The objective of this task is to perform spatio-temporal localization of the target object indicated by a textual query such as "the blue chair spinning on the left side of the screen" across all frames of a given video. This is a fundamental and challenging multimodal problem that requires not only recognizing an object but also understanding its attributes, spatial relationships, and dynamic context, including temporal changes and movements, thereby "grounding" language to visual information. While significant progress has been made in REC for static images, extending this task to videos introduces new dimensions of complexity. It demands that a model possess advanced spatio-temporal reasoning capabilities to capture appearance changes of the object over time and the dynamics of complex interactions between multiple objects.

Among the various forms of video, the importance of the egocentric recordings of the first-person views has grown rapidly in recent years [9, 16, 17, 33], driven by the anticipated proliferation of AR/VR devices, smart glasses [12], and collaborative robots [5]. In contrast to third-person videos that provide observers' viewpoint, egocentric videos provide the camera wearers' subjective experience. The camera motion directly reflects the wearer's focus of attention, while the appearance of their hands in the frame captures their physical inter-

actions with the environment, providing an unparalleled source of information for understanding user intent. As such, REC in the egocentric view is a foundational technology for the seamless integration of human-centric AI systems into our daily lives and professional workspaces [23,48]. However, research in this domain is still in its nascent stages, with a notable lack of benchmarks that reflect the complexity of the real world. The unique and challenging characteristics of the egocentric perspective—such as drastic viewpoint changes, frequent occlusions by the wearer's hands, and long-duration tasks—pose significant challenges to existing methods.

While previous research has made significant progress, existing benchmarks do not adequately capture the complex situations that humans encounter in the real world. Specifically, there are three main limitations or gaps: 1. Lack of Temporal Context: Many existing benchmarks are composed of short video clips, typically lasting from a few to tens of seconds. However, real-world tasks often require long-form memory and contextual understanding. 2. Simplified Scenarios: The tasks tend to be oversimplified, with the target object often being clearly visible for the majority of the video. In the real world, it is common for an object to appear only transiently or be situated among many confusing objects (distractors). 3. Neglect of Interaction: Most importantly, there is a significant lack of referring expressions that focus on dynamic and complex human-object interactions, such as a person manipulating an object or collaborating with others. This capability is essential for the next generation of AI assistants to be truly useful.

To overcome these challenges, in this work, we propose LongEgoRefer, a new egocentric video REC benchmark based on more realistic and challenging scenarios (see Figure 1). This benchmark was created by combining long-form videos from Ego4D with the precise object tracking annotations from EgoTracks [48]. The high-quality referring expressions were constructed using a hybrid approach: automatic generation by a Large Language Model (LLM) followed by careful verification and refinement by humans. As illustrated in the comparative analysis in Figure 1, our benchmark exhibits distinct characteristics that directly address the aforementioned limitations: 1. **Long-Form Videos**: By leveraging the large-scale egocentric video dataset Ego4D, we target long-form videos averaging 45 minutes in length to evaluate the capability for long-form contextual understanding (Figure 1a). 2. **Sparse Appearance**: We intentionally keep the appearance frequency of the referent low (increasing sparsity) to create a more realistic and difficult REC task (Figure 1b). 3. **Linguistic Complexity**: As such extensive footage inevitably contains numerous visually ambiguous scenes, we provide highly descriptive referring expressions to uniquely disambiguate the target from similar-looking temporal contexts. This results in significantly increased expression lengths compared to prior benchmarks (Figure 1c). 4. **Interaction-Centric**: The majority of our referring expressions are grounded in complex human-object interactions, demanding a high level of scene understanding (Figure 1, top).

To quantitatively demonstrate the difficulty of our proposed benchmark and to provide a starting point for the research community, we evaluate existing video REC models alongside a training-free baseline we construct. Instead of proposing a definitive end-to-end solution, our training-free baseline modularizes the task into two stages: temporal grounding, leveraging state-of-the-art Vision–Language Models (VLMs) to determine *when* an event occurs, and spatial grounding, utilizing Grounded SAM2 [39] to localize *where* the referred object appears. Our comprehensive experiments reveal a significant performance gap between open-source and closed-source models. Notably, while GPT-5 significantly outperforms all other models, the overall performance of both existing video REC models and our advanced VLM-based baselines remains modest, particularly under stricter evaluation metrics (e.g., mtIoU and mvIoU). This strictly indicates that the task of accurately localizing rare events and objects in long-form videos continues to be a significant and unresolved challenge even for current state-of-the-art models. The contributions of this research can be summarized as threefold:

- We introduce **LongEgoRefer**, a novel and challenging egocentric video REC benchmark. It specifically addresses the limitations of prior work by featuring hour-scale temporal contexts, extreme target sparsity, and complex interaction-centric referring expressions, thereby accurately reflecting realistic first-person visual experiences.
- We establish standardized baseline strategies to effectively evaluate models on extreme long-form videos. Specifically, we formulate two distinct approaches: (1) adapting existing short-clip video REC models via a sliding window inference strategy, and (2) constructing a training-free, two-stage pipeline that explicitly decouples the task into high-level temporal grounding (via VLMs) and low-level spatial tracking (via Grounded SAM2).
- We provide a comprehensive empirical analysis exposing the fundamental bottlenecks of current paradigms. We reveal that existing video REC models applied in a sliding-window manner lack global temporal awareness. Meanwhile, the two-stage pipeline suffers from critical error propagation, where inaccuracies in identifying when an event occurs strictly constrain localizing where it happens.

2 Related Work

2.1 Video-based REC Datasets

Video-based REC extends image-based grounding [21,31,34,40,60] to the spatio-temporal domain, aiming to localize target objects in videos based on natural language queries. Early video REC studies primarily adopted short, trimmed clips (3–10 seconds) and focused on detecting objects or events within brief frame sequences [22, 26, 42, 58, 65]. Concurrent work explored dynamic human activities and object interactions in third-person videos [14, 43, 44, 49, 65]. Recent research has moved toward egocentric viewpoints to capture everyday activities.

MeViS [11] scales referring video object segmentation to multi-event, motion-centric scenarios in third-person videos. RefEgo [23] introduces first-person REC over short clips lasting about 12 seconds on average, with temporal dependencies. EgoMask [27] benchmarks fine-grained spatio-temporal grounding in egocentric videos but covers segments of about 90 seconds with frequently appearing targets. However, these benchmarks are limited in temporal scale and do not address the challenge of localizing rare object occurrences across extended temporal contexts. LongEgoRefer advances this direction by introducing hour-scale first-person recordings with sparse target appearances, requiring models to maintain long-term memory while performing precise spatial grounding.

2.2 Egocentric Video Benchmarks

Egocentric video benchmarks are essential for understanding human activities from a first-person perspective, directly relevant to embodied AI and human-robot interaction. Large-scale wearable-camera datasets [9, 16, 29, 41] have captured daily activities with multi-modal data including RGB, depth, audio, and gaze. Recent benchmarks further incorporate multi-view synchronization and 3D scene geometry [17,32,33,51,52]. Building on these datasets, particularly Ego4D, language-conditioned egocentric understanding has emerged across tasks such as video summarization [7,67], video question answering [4,10,30], and video visual grounding [23,27,35]. While these works have made significant progress, existing grounding benchmarks focus on short, densely annotated clips where target objects appear frequently and predictably. In contrast, real-world egocentric videos contain sparse, irregular object occurrences distributed across hours of continuous recording. LongEgoRefer addresses this gap by requiring models to perform spatio-temporal grounding in hour-long videos where targets may appear only briefly and unpredictably, better reflecting realistic egocentric scenarios.

2.3 Long-form Video Understanding

Understanding long-form videos is critical for this work, as it requires models to integrate information across extended time spans rather than within isolated clips. Early video understanding focused on short, trimmed clips (10–100 seconds) with localized reasoning [20, 57, 61]. Subsequent datasets increased narrative complexity but remained limited to several minutes [24, 37, 55]. Recent benchmarks extend to long-form comprehension spanning tens of minutes to hours [7, 13, 30, 46, 47, 54, 66], emphasizing multi-event reasoning and cross-scene consistency. However, these works primarily address event-level temporal reasoning without spatial grounding. In contrast, LongEgoRefer requires models to localize rare objects across extended temporal contexts while maintaining precise frame-level spatial grounding. This setting challenges models to combine long-term memory for tracking rare occurrences with accurate language–visual alignment for fine-grained localization in long-form videos.

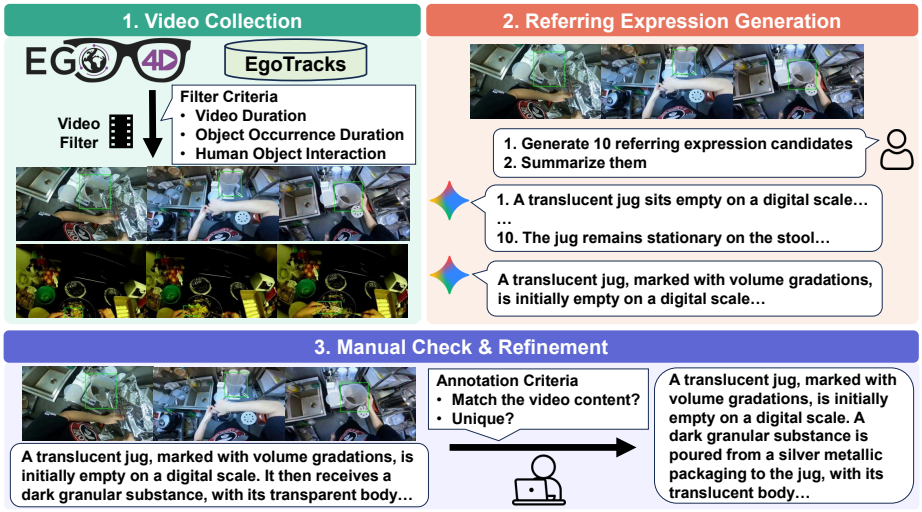

Fig. 2: An overview of our benchmark construction pipeline.

3 Benchmark

3.1 Task

We address the task of *Spatio–Temporal Grounding of Object Occurrences* in long-form egocentric videos. Given an untrimmed video V of length T_V and a natural language query Q, the model must localize a single, continuous appearance of the referred object within the untrimmed video. The prediction is represented as a tuple $(t_{\text{start}}, t_{\text{end}}, \mathcal{B})$, where $[t_{\text{start}}, t_{\text{end}}]$ denotes the temporal boundaries satisfying $0 \leq t_{\text{start}} \leq t_{\text{end}} \leq T_V$, and $\mathcal{B}$ is the sequence of bounding boxes corresponding to that duration. The key challenge is to find a short, language-relevant moment within hours of egocentric footage and precisely localize the target object throughout that interval, jointly testing temporal localization and spatial grounding under realistic, first-person conditions.

3.2 Benchmark Creation

Dataset and annotation preparation We build our benchmark upon the annotations from the EgoTracks dataset [48], which provides long-form object tracking annotations on the large-scale egocentric video corpus Ego4D [16]. EgoTracks contains per-frame bounding boxes for each object occurrence, along with instance-level attribute information. While EgoTracks primarily consists of six-minute clips extracted from the original Ego4D videos, our benchmark instead utilizes the *full-length* Ego4D videos to evaluate spatio–temporal grounding under truly long-form conditions. Although six-minute clips are longer than those in prior grounding datasets, untrimmed hour-scale videos are crucial to evaluate robustness to rare events and long-term temporal dependencies in realistic

scenarios. Importantly, EgoTracks retains timestamp metadata linking each clip to its position in the original video, allowing us to accurately remap existing annotations back to the full-length Ego4D sequences.

Annotation Pipeline To ensure the quality, diversity, and uniqueness of our referring expressions, we designed the semi-automated annotation pipeline illustrated in Figure 2, which consists of a data selection stage and a two-stage language generation process that guarantees each expression does not match other parts of the long video.
(i) Data filtering. To capture realistic long-term scenarios, we first perform strict filtering on the Ego4D dataset. We exclude all videos shorter than 20 minutes to ensure sufficient temporal context for long-form reasoning. To avoid trivial segments, we retain only object tracks lasting at least 10 seconds where the camera wearer actively manipulates the object (annotated as `is_active` = `True` in EgoTracks). This criterion prioritizes dynamic, interaction-rich episodes that demand temporal grounding.
(ii) Referring expression generation. For each filtered clip, we generate referring expressions using Gemini 2.5 Flash through a two-stage process. In the first stage, the model is prompted to produce 10 detailed captions describing the target object's interactions, state changes, and spatial relations. In the second stage, these captions are aggregated and refined into a single, natural referring expression that captures the temporal context. This step removes redundancy and produces a refined, cohesive referring expression that encapsulates the temporal context.

Human annotation All generated expressions undergo rigorous human verification to ensure their quality and correctness. Because automatically generated referring expressions can be semantically inconsistent or temporally ambiguous (that is, applicable to multiple moments in the video), we adopt a four-stage human curation pipeline conducted by expert annotators. As a result of this meticulous process, our annotators substantially revised 95.39% of the Gemini-initialized descriptions to precisely match the video content and ensure referential exclusivity. This extensive human intervention effectively mitigated potential model-induced biases, transforming the initial LLM-generated seeds into high-quality, human-curated grounding queries that provide a fair evaluation platform for all VLM families. Starting from 2,000 candidate triplets consisting of an Ego4D video, an object occurrence with bounding boxes, and a generated referring expression, we perform the following steps:
(i) Validating and filtering. We first check the video segment and temporal boundaries of each occurrence. Samples are removed if (a) the corresponding segment is blank or corrupted, or (b) the object remains visible before or after the annotated timestamps, which makes precise temporal grounding unreliable.
(ii) Correcting expressions. For each valid sample, an annotator verifies whether the referring expression accurately describes the target object and its associated actions. If inaccuracies are found, the expression is rewritten to be factually and semantically correct.

Benchmark	HOI	Length (s)	Appear (%)	#Videos	#Ref	#Words
Exocentric Benchmark						
Lingual OTB99 [26]	✗	19.9	∼100	99	99	5.5
Lingual ImageNet Videos [26]	✗	9.1	∼100	100	100	5.6
ReferDAVIS-16 [22]	✗	2.9	∼100	50	50	5.9
ReferDAVIS-17 [22]	✗	2.9	∼100	120	120	4.7
Refer-Youtube-VOS [42]	✗	4.5	∼100	3,978	15,009	10.0
MeViS [11]	✗	13.2	∼100	2,006	28,570	8.5
Egocentric Benchmark						
EgoMask [27]	✗	91.8	60.3	315	700	15.0
$\text{RefEgo}_{\text{test}}$ [23]	✗	11.9	75.5	537	1,317	17.5
LongEgoRefer (Ours)	✓	2,714.9	1.3	433	1,498	63.5

Table 1: Comparison between LongEgoRefer and existing video REC datasets and benchmarks. "Appear" is the appearance rate of the referred object in frames. "#Ref" is the number of referring expressions and "#Words" is average word length.

(iii) Disambiguating references. The annotator then ensures that the refined expression uniquely identifies the target moment. If the same description fits other segments in the video, additional visual or temporal cues (such as object attributes, spatial relations, or nearby actions) are added until the expression becomes exclusive to the intended instance.
(iv) Quality assurance. A second annotator independently reviews the results from the previous stages, returning 52.20% of them for re-annotation until they fully pass the strict quality check. Ten percent of all annotations are selected for an additional verification round to ensure consistency and final quality.

Through this rigorous pipeline, 25.1% of the initial candidates were filtered out, resulting in a final dataset of 1,498 high-quality curated samples.

3.3 Statistics

The proposed LongEgoRefer benchmark provides a more realistic and challenging setting than previous video REC benchmarks. Table 1 compares LongEgoRefer with existing video referring expression and grounding benchmarks. Our proposed LongEgoRefer substantially increases both temporal and linguistic complexity compared to existing video grounding benchmarks. It contains 433 egocentric videos with average length of 2,714.9 seconds (approximately 45 minutes) and 1,498 referring expressions with an average of 63.5 words per query, which are significantly longer and more descriptive than those in prior datasets. While most existing benchmarks capture short segments where the referred object remains visible throughout the clip (with an appearance rate of 60.3% or higher), the target objects in LongEgoRefer appear in only 1.3% of frames, requiring models to perform long-range temporal search, contextual reasoning, and memory maintenance over extended sequences. Figure 1 highlights clear differences between RefEgo and LongEgoRefer. RefEgo contains short clips with frequent object appearances, whereas LongEgoRefer consists of much longer videos with sparse targets and longer descriptions. Furthermore, LongEgoRefer explicitly

annotates human–object interactions (HOI), creating a challenging setting for spatio-temporal grounding in real-world, continuous egocentric videos. These differences make LongEgoRefer a comprehensive benchmark for evaluating models' ability to perform robust, fine-grained spatio-temporal grounding over extended egocentric video sequences, bridging research and real-world applications.

4 Baseline

To comprehensively evaluate the difficulty of the LongEgoRefer benchmark and provide a starting point for future research, we construct and evaluate two types of baselines: existing end-to-end video REC models adapted for long videos, and a training-free two-stage pipeline leveraging modern visual foundation models.

4.1 Video REC Models

Existing video REC models are predominantly designed and trained on short, trimmed clips (typically spanning a few seconds). Consequently, they are fundamentally constrained by memory limitations and context window sizes, making them infeasible to process hour-long videos natively.

To evaluate these conventional models on the extensive sequences of LongEgoRefer, we adopt a sliding window inference strategy. Specifically, we partition the untrimmed long video into a sequence of short temporal windows. We then independently feed each window, along with the referring expression, into the end-to-end video REC model. The model generates a sequence of bounding boxes and an associated confidence score for each window.

The aggregation strategy for the final prediction depends on the output capabilities of the specific model. For models capable of generating confidence scores for their predictions, we select the temporal window and its corresponding spatio-temporal track that yields the highest confidence score across the entire video. For models that do not provide such evaluation metrics, we instead aggregate the results by temporally concatenating the predicted bounding boxes from all windows to form a single continuous sequence. This sliding window approach serves as a straightforward baseline to gauge how well current short-clip REC models generalize to extreme long-form search scenarios.

4.2 Training-free Two-stage Pipeline

Spatio-temporal grounding in long-form videos poses significant computational challenges for end-to-end models. To address this without extensive retraining, we construct a hierarchical pipeline that decouples the task into temporal and spatial reasoning. This design leverages a VLM for global temporal reasoning ("when") and Grounded SAM2 [39] for local spatial localization ("where"). We adopt Grounded SAM2 for the spatial stage because of its SoTA performance on EgoMask [27].

Model	Temporal			Spatio-Temporal				
	tIoU@0.1	tIoU@0.5	mtIoU	vIoU@0.1	vIoU@0.5	mvIoU	mSTIoU	mIoU+n
Video REC models								
VideoLISA [3]	0.45	0.07	0.15	0.13	0.00	0.37	0.34	4.90
Sa2VA [62]	0.60	0.00	1.33	2.99	0.00	1.72	2.03	62.75
Grounded SAM2 [39]	0.47	0.00	1.28	0.33	0.00	0.67	0.51	26.99
SAM3 [6]	0.56	0.09	0.19	0.45	0.07	0.15	0.18	**98.82**
Open-source VLMs								
InternVideo 2.5 Chat 8B [53]	1.34	0.27	0.48	0.46	0.07	0.18	0.21	96.34
MiMo-VL 7B RL [56]	1.13	0.20	0.42	0.33	0.00	0.13	0.15	92.65
TimeChat [38]	2.00	0.20	0.79	0.73	0.00	0.24	0.28	93.31
LITA [18]	2.60	0.06	1.14	1.46	0.06	0.50	0.59	85.99
Video-Chat Flash 7B [25]	3.07	0.47	1.07	1.20	0.13	0.37	0.45	82.45
mPLUG-Owl3 7B [59]	2.60	0.33	0.89	1.27	0.20	0.45	0.51	88.22
LLaVA-OneVision 1.5 8B [1]	1.20	0.33	0.57	0.67	0.07	0.19	0.20	80.58
VideoLLaMA3 7B [63]	3.34	0.20	1.20	1.54	0.07	0.49	0.63	52.85
LongVA 7B DPO [64]	4.21	0.53	1.30	1.54	0.20	0.51	0.61	82.51
LongVILA R1 7B [8]	2.27	0.40	0.91	1.00	0.13	0.33	0.45	66.67
InternVL 3.5 8B [50]	4.34	0.07	1.50	2.27	0.07	0.66	0.84	15.98
Qwen3-VL 8B [2]	4.74	0.93	1.73	1.94	0.47	0.83	0.93	90.28
InternVL 3.5 38B [50]	6.34	0.40	1.98	2.87	0.33	0.94	1.07	78.54
Qwen3-VL 32B [2]	4.87	1.07	1.81	2.20	0.47	0.84	0.97	89.83
Closed-source VLMs								
Gemini 2.0 Flash [15]	5.14	1.87	2.29	2.73	0.93	1.16	1.33	97.43
Gemini 2.5 Flash [15]	10.08	3.47	4.49	5.14	1.66	2.21	2.62	95.26
Gemini 2.5 Pro [15]	23.30	8.68	10.73	12.01	4.13	5.49	6.42	95.76
GPT-4o [19]	19.69	6.54	8.71	10.68	3.47	4.49	5.11	94.55
GPT-5 [45]	**37.31**	**12.55**	**16.19**	**17.48**	**6.20**	**7.89**	**8.76**	95.77

Table 2: Experimental results on our LongEgoRefer benchmark. We evaluate a comprehensive set of open- and closed-source VLMs.

Specifically, we first employ a pre-trained VLM to identify the temporal interval that matches the query. This coarse estimate is then refined into a precise trajectory using Grounded SAM2. We extract the middle frame of the identified clip and use Grounding DINO [28] to detect the target object and generate an initial bounding box. This box serves as a visual prompt for SAM2 [36], which tracks the object across the segment to produce the final spatio-temporal track. This modular approach provides a clear reference point for the LongEgoRefer benchmark.

5 Experiments

5.1 Experimental Settings

Metrics For the evaluation on LongEgoRefer, we adopt four metrics. tIoU and vIoU are commonly used metrics for video REC. Following RefEgo, we also employ STIoU and IoU+n.

Implementation Details We evaluate 4 video REC models, and 14 open-source and 5 closed-source VLMs on LongEgoRefer. The video REC models are VideoLISA [3], Sa2VA [62], Grounded SAM2 [39], and SAM3 [6]. The open-source

models are InternVideo 2.5 Chat [53], MiMo-VL [56], TimeChat [38], LITA [18], Video-Chat [25], mPLUG-Owl3 [59], LLaVA-OneVision [1], Video-LLaMA3 [63], LongVA [64], LongVILA R1 [8], InternVL 3.5 [50], and Qwen3-VL [2]. The closed-source models include Gemini-family (2.0 Flash, 2.5 Flash, and 2.5 Pro) [15] and GPT-family (GPT-4o [19] and GPT-5 [45]). See Appendix for detailed information.

5.2 Main Results

Table 2 presents the quantitative evaluation results of video REC models and leading open-source and closed-source VLMs on our proposed LongEgoRefer benchmark. A key finding from our experiments is that even state-of-the-art models struggle with referring expression comprehension in long-form egocentric videos. The generally low scores across temporal and spatio-temporal metrics highlight the inherent difficulty of accurately localizing a specific clip, often lasting only tens of seconds, within a video that spans tens of minutes to an hour, based solely on referring expressions.

Notably, the poor performance of video REC models highlights that existing approaches are technically incapable of handling LongEgoRefer directly. The sliding window strategy, while a natural baseline, fails because it lacks the global temporal awareness needed to filter out irrelevant segments across an entire hour, leading to high false-positive matches. This underscores the necessity of our two-stage pipeline: by decomposing the task into temporal and spatial grounding, we leverage the VLM's extensive context window to prune the vast search space. This "temporal filtering" allows the video REC model to focus its fine-grained spatial localization on a manageable, contextually relevant clip, which would otherwise be buried in hours of irrelevant data.

Comparisons across model categories reveal a substantial performance gap between open-source and closed-source models. Even relatively large and high-performing open-source models, such as InternVL and Qwen3-VL, remain in the single digits for tIoU@0.1. Furthermore, it is noteworthy that models specifically designed for long video understanding, such as LongVILA, also show limited effectiveness on our benchmark. In contrast, the closed-source models achieve significantly better performance. Notably, GPT-5 sets a new state of the art, with a tIoU@0.1 of 37.31 and an mtIoU of 16.19, outperforming all other models by a wide margin. Importantly, the fact that GPT-5 consistently outperforms the Gemini series, which was used to generate the initial referring expressions, demonstrates that the benchmark does not suffer from a Gemini-specific bias. This result, coupled with the rigorous human curation process where 95.39% of initial descriptions were substantially revised (see Section 3.2), confirms that LongEgoRefer evaluates objective spatio-temporal reasoning rather than rewarding model-specific linguistic patterns. This advantage extends to spatio-temporal grounding. While most open-source models struggle to produce accurate object tracks (reflected in mSTIoU scores below 1.0), the closed-source models demonstrate stronger, though still developing, capabilities. GPT-5 again leads, achieving an mSTIoU of 8.76 and an mvIoU of 7.89. Within the Gemini family, Gemini

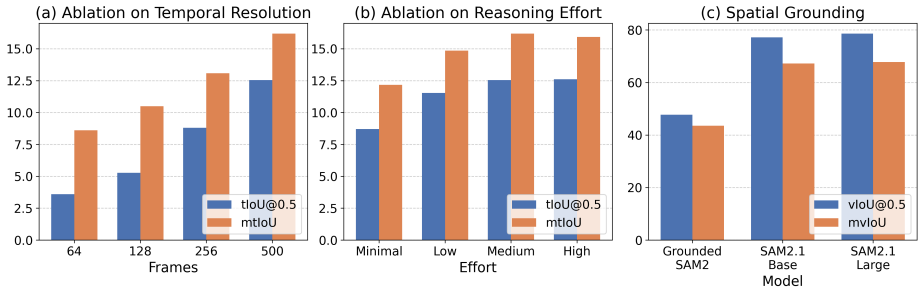

Fig. 3: Analysis of temporal resolution, reasoning effort, and spatial grounding.

2.5 Pro performed best, and 2.5 Flash followed, suggesting that both the model scale and architectural sophistication contribute directly to the performance improvements on complex long-form context understanding.

Despite achieving the best results, the absolute performance of GPT-5 remains relatively low, indicating substantial room for improvement. Temporal localization accuracy appears to be the primary limiting factor, constraining subsequent spatio-temporal grounding metrics. Overall, these results empirically demonstrate that LongEgoRefer presents a highly challenging problem, underscoring the need for novel approaches capable of efficiently processing and deeply understanding long-form video content.

5.3 Analysis

Effect of Temporal Resolution Figure 3a investigates the impact of temporal sampling density on GPT-5 by reducing the maximum frames from the default 500 (constrained by API limits). We observe a significant performance decay as the temporal resolution decreases. Specifically, tIoU@0.5 drops sharply from 12.55 to 3.60 when the input is limited to 64 frames. This suggests that sparse sampling leads to the loss of fine-grained visual cues essential for precise temporal localization in long-form videos. These results confirm that high temporal resolution is critical for accurate grounding and that GPT-5 effectively leverages dense 500-frame contexts to model complex temporal dependencies.

Effect of Reasoning Effort Figure 3b investigates the impact of reasoning effort levels on GPT-5. Specifically, "Minimal" corresponds to zero reasoning effort, whereas "High" represents the maximum available reasoning capacity. The results demonstrate that increasing reasoning effort yields substantial performance gains, with tIoU@0.5 improving from 8.7 to 12.61. Notably, we observe that performance largely saturates at the "Medium" level, which serves as the API's default setting. This suggests that while more intensive reasoning generally benefits temporal localization, there may be a point of diminishing returns or increased noise.

Model	Temporal			Spatio-Temporal				
	tIoU@0.1	tIoU@0.5	mtIoU	vIoU@0.1	vIoU@0.5	mvIoU	mSTIoU	mIoU+n
Gemini 2.0 Flash	39.38	14.35	17.70	19.49	6.40	8.44	9.78	**96.95**
Gemini 2.5 Flash	45.19	15.28	20.36	22.76	**8.21**	10.21	11.83	95.89
Gemini 2.5 Pro	**49.93**	**19.35**	**22.97**	**25.43**	8.07	**10.88**	**12.63**	96.44

Table 3: Experimental results on EgoTracks videos.

Spatial Grounding We further investigate the importance of temporal grounding. We evaluate spatial grounding alone with the annotated (oracle) and Grounded-SAM2-based temporal segmentation for the object appearance of each video. With the oracle temporal segmentation, we evaluate the segmentation model of SAM2.1 Base and Large. We use the ground-truth bounding box in the first frame for the prompts of SAM2.1. Figure 3c presents the results in terms of vIoU. Compared to the results in Table 2, the scores have improved dramatically with the oracle temporal segmentation. For example, while GPT-5 achieves an mvIoU of 7.89 when using its own temporal grounding results, Grounded SAM2 attains 43.54 (+35.65). This demonstrates that the accuracy of temporal grounding has a substantial impact on the overall performance of spatio-temporal grounding. Furthermore, when comparing Grounded SAM2 with SAM2.1 Base and SAM2.1 Large, mvIoU increases to 67.21 (+23.67) and 67.82 (+24.28), respectively. These results indicate that the precision of the bounding box used as the prompt for SAM2 significantly contributes to the performance of spatial grounding.

Results on EgoTracks Videos To better understand the challenges of long-form videos in video REC, we evaluate Gemini on LongEgoRefer using the EgoTracks dataset instead of the original Ego4D videos. EgoTracks is constructed by trimming the Ego4D videos to an average length of six minutes. Although shorter than the original Ego4D videos (45 minutes on average), EgoTracks still consists of considerably longer videos than those in RefEgo or EgoMask. The results are summarized in Table 3. Compared to the results on the original long-form Ego4D videos, the performance improves substantially. For instance, Gemini 2.5 Pro achieves a gain of +12.24 mtIoU and +5.39 mvIoU. These results indicate that while long-form video REC is an important and practical problem, it remains highly challenging. We hypothesize that the main factors behind the performance degradation on long videos are the repeated occurrence of visually similar events and the decreased frequency of target object appearances.

5.4 Qualitative Results

We present prediction examples of InternVL 3.5 38B, GPT-5, and ground truth in Figure 4. In the example (a), both InternVL and GPT predict the correct bounding boxes. In the example (b), GPT predicts the correct bounding boxes, but InternVL predicts the wrong temporal duration and bounding boxes. In the

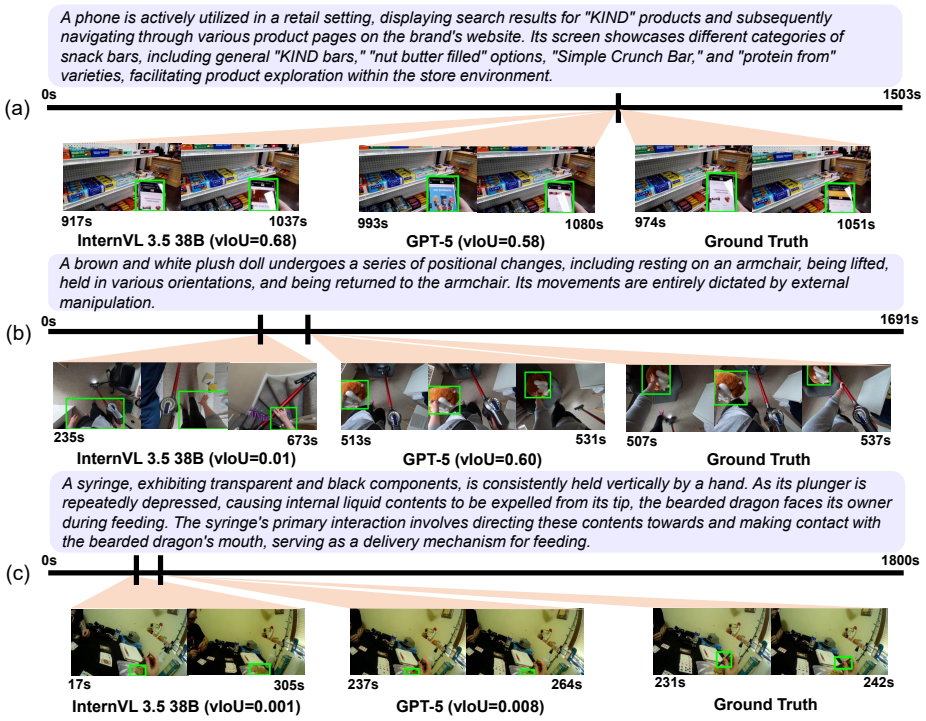

Fig. 4: Qualitative results.

example (c), both InternVL and GPT predict the wrong bounding boxes, even though the temporal duration of GPT is close to the ground truth.

6 Conclusion

In this work, we introduced **LongEgoRefer**, a new benchmark for egocentric video referring expression comprehension that addresses key limitations of existing benchmarks. It substantially expands the temporal scale (averaging 45 minutes per video) and introduces extreme target sparsity, where referents are visible for only 1.3% of the video, requiring models to identify brief and infrequent visual occurrences within long, continuous streams. Furthermore, LongEgoRefer is the first egocentric benchmark explicitly centered on complex human-object interactions (HOI), accompanied by long, descriptive expressions averaging over 63 words, which demand advanced spatio-temporal reasoning and language grounding. Our experiments show that these characteristics present significant challenges to current vision-language models. Even the strong training-free baselines we constructed, which combine state-of-the-art VLMs with the powerful spatial tracker, achieve limited accuracy. LongEgoRefer thus establishes a new standard for real-world complexity and provides a foundation for developing models with

robust long-term memory, interaction reasoning, and the ability to ground complex language in sparse and cluttered visual environments, advancing research toward more capable human-centric AI systems.

This study focuses on constructing a rigorous benchmark for spatio-temporal grounding in long-form egocentric videos. While LongEgoRefer provides a critical testbed, developing a comprehensive, large-scale training dataset for videos spanning tens of minutes remains an important direction for future work. For practical applications such as human-robot interaction in daily-life scenarios, extending the benchmark to multi-hour or day-long continuous videos will be indispensable.

Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR22P8, JST K Program, Grant Number JPMJKP25V2, JST CRONOS, Grant Number JPMJCS24K6, and JST BOOST, Grant Number JPMJBY24C5. This work used the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Liu, L., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. In: Adv. Neural Inform. Process. Syst. pp. 6833–6859 (2024)
4. Bärmann, L., Waibel, A.: Where did i leave my keys? — episodic-memory-based question answering on egocentric videos. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 1560–1568 (2022)
5. Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al.: Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 (2025)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
7. Chandrasegaran, K., Gupta, A., Hadzic, L.M., Kota, T., He, J., Eyzaguirre, C., Durante, Z., Li, M., Wu, J., Li, F.F.: Hourvideo: 1-hour video-language understanding. In: Adv. Neural Inform. Process. Syst. pp. 53168–53197 (2024)
8. Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., He, E., Yin, H., Molchanov, P., Kautz, J., Fan, L., Zhu, Y., Lu, Y., Han, S.: Longvila: Scaling long-context visual language models for long videos. In: Int. Conf. Learn. Represent. (2025)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: Eur. Conf. Comput. Vis. pp. 720–736 (2018)

10. Di, S., Xie, W.: Grounded question-answering in long egocentric videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12934–12943 (2024)
11. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: Int. Conf. Comput. Vis. pp. 2694–2703 (2023)
12. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., et al.: Project aria: A new tool for egocentric multi-modal ai research. arXiv preprint arXiv:2308.13561 (2023)
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 24108–24118 (2025)
14. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5958–5966 (2018)
15. Gemini Team: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18995–19012 (2022)
17. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19383–19400 (2024)
18. Huang, D.A., Liao, S., Radhakrishnan, S., Yin, H., Molchanov, P., Yu, Z., Kautz, J.: Lita: Language instructed temporal-localization assistant. In: Eur. Conf. Comput. Vis. pp. 202–218 (2024)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2758–2766 (2017)
21. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Conference on Empirical Methods in Natural Language Processing. pp. 787–798 (2014)
22. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: Asian Conf. Comput. Vis. pp. 123–141 (2018)
23. Kurita, S., Katsura, N., Onami, E.: Refego: Referring expression comprehension dataset from first-person perception of ego4d. In: Int. Conf. Comput. Vis. pp. 15214–15224 (2023)
24. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22195–22206 (2024)
25. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., Qiao, Y., Wang, Y., Wang, L.: Videochat-flash: Hierarchical compression for long-context video modeling. In: Int. Conf. Learn. Represent. (2026)

26. Li, Z., Tao, R., Gavves, E., Snoek, C.G.M., Smeulders, A.W.M.: Tracking by natural language specification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6495–6503 (2017)
27. Liang, S., Zhong, Y., Hu, Z.Y., Tao, Y., Wang, L.: Fine-grained spatiotemporal grounding on egocentric videos. In: Int. Conf. Comput. Vis. pp. 9385–9395 (2025)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Eur. Conf. Comput. Vis. pp. 38–55 (2024)
29. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: HOI4D: A 4d egocentric dataset for category-level human-object interaction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21013–21022 (2022)
30. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: Adv. Neural Inform. Process. Syst. pp. 46212–46244 (2023)
31. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11–20 (2016)
32. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: Int. Conf. Comput. Vis. pp. 20133–20143 (2023)
33. Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K.K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., Chalk, J., Zhu, Z., Guerrier, R., Abdelazim, F., Zhu, B., Moltisanti, D., Wray, M., Doughty, H., Damen, D.: HD-EPIC: A highly-detailed egocentric video dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23901–23913 (2025)
34. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Int. Conf. Comput. Vis. pp. 2641–2649 (2015)
35. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: Naq: Leveraging narrations as queries to supervise episodic memory. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6694–6703 (2023)
36. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. In: Int. Conf. Learn. Represent. (2025)
37. Rawal, R., Saifullah, K., Basri, R., Jacobs, D., Somepalli, G., Goldstein, T.: Cinepile: A long video question answering dataset and benchmark. arXiv preprint arXiv:2405.08813 (2024)
38. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: Timechat: A time-sensitive multimodal large language model for long video understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14313–14323 (2024)
39. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024)
40. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Eur. Conf. Comput. Vis. pp. 817–834 (2016)
41. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21096–21106 (2022)

42. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: Eur. Conf. Comput. Vis. pp. 208–223 (2020)
43. Shang, X., Li, Y., Xiao, J., Ji, W., Chua, T.S.: Video visual relation detection via iterative inference. In: ACM Int. Conf. Multimedia. pp. 3654–3663 (2021)
44. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: ACM Int. Conf. Multimedia. pp. 1300–1308 (2017)
45. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
46. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., Lu, Y., Hwang, J.N., Wang, G.: Moviechat: From dense token to sparse memory for long video understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 18221–18232 (2024)
47. Tan, X., Luo, Y., Ye, Y., Liu, F., Cai, Z.: Allvb: All-in-one long video understanding benchmark. In: AAAI. pp. 7211–7219 (2025)
48. Tang, H., Liang, K.J., Grauman, K., Feiszli, M., Wang, W.: Egotracks: A long-term egocentric visual object tracking dataset. In: Adv. Neural Inform. Process. Syst. pp. 75716–75739 (2023)
49. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. IEEE Trans. Circuit Syst. Video Technol. **32**(12), 8238–8249 (2021)
50. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
51. Wang, X., Zhao, K., Liu, F., Wang, J., Zhao, G., Bao, X., Zhu, Z., Zhang, Y., Wang, X.: Egovid-5m: A large-scale video-action dataset for egocentric video generation. In: Adv. Neural Inform. Process. Syst. (2025)
52. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., Joshi, N., Pollefeys, M.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Int. Conf. Comput. Vis. pp. 20270–20281 (2023)
53. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., Dou, M., Chen, K., Wang, W., Qiao, Y., Wang, Y., Wang, L.: InternVideo2.5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
54. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: Adv. Neural Inform. Process. Syst. pp. 28828–28857 (2024)
55. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9777–9786 (2021)
56. Xiaomi LLM-Core Team: Mimo-vl technical report. arXiv preprint arXiv:2506.03569 (2025)
57. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM Int. Conf. Multimedia. pp. 1645–1653 (2017)
58. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.S.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Eur. Conf. Comput. Vis. pp. 585–601 (2018)

59. Ye, J., Xu, H., Liu, H., Hu, A., Yan, M., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In: Int. Conf. Learn. Represent. (2025)
60. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Eur. Conf. Comput. Vis. pp. 69–85 (2016)
61. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. pp. 9127–9134 (2019)
62. Yuan, H., Li, X., Zhang, T., Sun, Y., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
63. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
64. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
65. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10668–10677 (2020)
66. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: Benchmarking multi-task long video understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13691–13701 (2025)
67. Zhou, W., Cao, K., Zheng, H., Liu, Y., Zheng, X., Liu, M., Kristensson, P.O., Mayol-Cuevas, W.W., Zhang, F., Lin, W., Shen, J.: X-LeBench: A benchmark for extremely long egocentric video understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 15206–15222 (2025)