EgoSAT: A Comprehensive Benchmark of Egocentric Streaming Interaction Understanding

Yijia Lei[1], Jinzhao Li[1†], Yichi Zhang[1], Jiacheng Hua[1], Yin Li[2], and Miao Liu[1‡]

[1] College of AI, Tsinghua University
[2] University of Wisconsin–Madison
{leiyj23,lijinzha22}@mails.tsinghua.edu.cn, miaoliu@mail.tsinghua.edu.cn

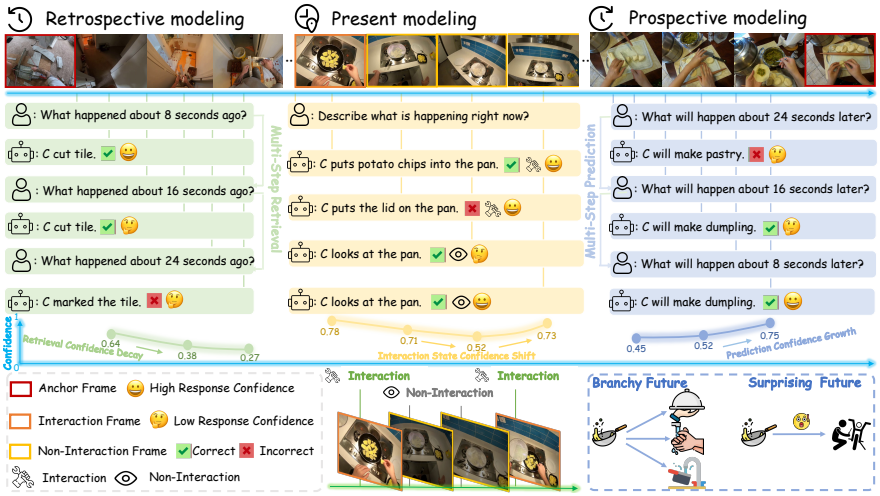

Fig. 1: EgoSAT presents a unified formulation that brings together several conventional vision–language tasks, *e.g.*, video question answering, online video narration, and activity anticipation, within a single streaming setting. In doing so, it provides the first comprehensive benchmark for evaluating the ability of modern vision–language models to reason about the *past, present, and future* under streaming observations.

Abstract. We introduce **EgoSAT**, the first comprehensive benchmark for egocentric video reasoning in streaming settings, designed to evaluate the capabilities of modern vision–language models (VLMs). The benchmark targets streaming interaction understanding, where video frames arrive sequentially and models must continuously interpret evolving visual context. EgoSAT unifies several previously distinct tasks within a single streaming framework. In this formulation, queries about completed events correspond to retrospective reasoning, queries about ongoing activities require online understanding, and queries about future actions involve prospective anticipation. This unified setting requires models to reason about the past, present, and future while operating under the constraint that only previously observed frames are available. EgoSAT

† Project Lead. ‡ Corresponding author.

contains 1,997 unique videos spanning 165 hours of egocentric footage and around 4,800 high-quality question–answer pairs, carefully designed to probe reasoning across varying temporal contexts. Using this benchmark, we evaluate a diverse set of both open-weight and closed-weight VLMs, providing a systematic assessment of their ability for streaming interaction understanding. By distinguishing answerability and conducting diagnostics on confidence of models, we find existing models not only struggle with prospective and retrospective modeling, but also exhibit severe mis-calibration: confidence often fails to track inherent answerability, leading to dangerous "confidently wrong" behaviors. Project page: https://leiyj23.github.io/EgoSAT/

Keywords: Egocentric Vision · Temporal Reasoning · Multimodal LLMs

1 Introduction

Advances in wearable cameras and edge computing, together with recent breakthroughs in vision language models (VLMs), have ushered in a new generation of AI systems that provide context-aware assistance to camera wearers through natural language interfaces, often combined with voice commands and gesture-based controls. A key characteristic of this setting is *streaming processing*: the AI assistant must continuously perceive, comprehend, and retain both the video captured by the device and queries issued by the user, in order to reason about past events (*e.g.*, video question answering), describe ongoing activities (*e.g.*, online video narration), and predict future intent (*e.g.*, activity anticipation).

Traditionally, these problems have been studied in silos, with dedicated methods and benchmarks designed for individual tasks. However, evaluating VLMs on isolated tasks provides only a partial assessment of their capabilities in realistic streaming scenarios, where perception and reasoning about past, present, and future events is inherently intertwined. In practice, an AI assistant must seamlessly integrate these capabilities within a unified, temporally evolving context.

Our key insight is that *many previously distinct tasks can be naturally reformulated within a single streaming framework*. In this setting, video frames arrive sequentially as a continuous stream, and user queries in text form may occur at arbitrary time points, and models are restricted to using only the portion of the video stream observed up to the time of the query. As shown in Figure 1, a query about a completed event (*i.e.*, **retrospective**) corresponds to standard video question answering (QA); a query about an ongoing event (*i.e.*, **present**) instantiates online video QA; and a query about a future event (*i.e.*, **prospective**) falls under event anticipation.

Beyond producing accurate answers, they must infer the temporal context of each query, and determine whether sufficient evidence has been observed. Moreover, they must continuously update their confidence as new evidence accumulates, and appropriately calibrate their uncertainty in light of multiple plausible future outcomes. This unified formulation enables a more realistic and holistic evaluation of streaming video understanding in the context of AI assistants

To this end, we introduce **EgoSAT**, the first benchmark for Egocentric Streaming interaction understanding, designed to evaluate the temporal reasoning capabilities of modern VLMs. We leveraged Ego4D, a scenario-rich and interaction-dense egocentric video corpus with careully curated temporal annotations.

Leveraging EgoSAT, we conduct extensive evaluation of both closed- and open- weight MLLMs under a strict online protocol. Across all settings, we find that prospective anticipation and retrospective retrieval remain challenging even for frontier models, while online streaming baselines further lag behind offline counterparts due to irreversible input-side compression. Finally, our answerability-aware confidence diagnostics reveal that model confidence often fails to track factual answerability, frequently staying high on inherently uncertain or incorrect predictions.

Our main contributions are summarized as follows.

1. We present **EgoSAT, a comprehensive benchmark** for egocentric streaming interaction understanding, designed to evaluate the temporal reasoning capabilities of modern VLMs.
2. Our key **technical innovations** are (1) *a unified streaming reasoning formulation* integrates several conventional vision-language tasks and requires models to reason about the past, present and future with streaming observations; and (2) *a principled quantification of answerability and its corresponding benchmark design* for future event prediction, enabling fair and systematic evaluation of VLMs.
3. Leveraging EgoSAT, we conduct **a systematic evaluation** of a diverse set of both open-weight and closed-weight vision–language models. Our empirical results reveals that (1) *prospective anticipation and retrospective retrieval remain challenging across model families*, and online streaming models further lag behind offline counterparts; (2) *confidence is often poorly calibrated to factual answerability*, with several models remaining confidently wrong, motivating answerability-aware diagnostics beyond accuracy.

2 Related Work

Streaming Visual Language Models. Streaming visual language models operate under an online prefix constraint where each query can only use frames observed so far. This setting requires low latency responses, continual state updates, and sustained reasoning over long streams. Vinci presents a real time embodied assistant based on egocentric vision language models [12], and Dispider introduces a disentangled perception, decision, and reaction framework for active real time interaction [19]. Streaming long video understanding with large language models(LLMs) proposes a streaming framework that processes videos segment by segment with memory propagation [20], while STREAMCHAT studies streaming video understanding with multiround interaction and memory enhanced knowledge [30]. Streaming VLM targets real time understanding for infinite video streams and aligns training with streaming inference through com-

pact memory management [31]. VideoLLM-online introduces the LIVE framework with efficient visual token encoding and decoding for online interaction [6], and TimeChat-Online proposes to prune redundant visual tokens in streaming videos [35]. LiveCC builds a streaming speech transcription pipeline for temporally aligned training and provides a benchmark for streaming video language models [7].

These works focus on system design, memory management, and efficiency for continuous streams. In egocentric settings, critical evidence is often localized around hands, objects, and gaze, and evidence can decay when background tokens dominate the budget. Our approach targets this interaction blind gap by prioritizing Region-of-Interest (ROI)-aware token budgeting and by training confidence to track evidence accumulation and decay for real time decisions.

Efficient Token Compression for Long Video Understanding. A large body of work improves efficiency by compressing or selecting visual tokens for long video and multimodal understanding. LongVU explores spatiotemporal adaptive compression for long video understanding [24], and PVC proposes progressive visual token compression across frames [33]. TESTA aggregates temporal and spatial tokens to condense video semantics [21], and VideoChat-Flash introduces hierarchical compression for long context video modeling [15]. DeCo analyzes visual projector design and decouples compression from semantic abstraction [34]. TokenLearner learns a compact set of informative tokens [22], and VQToken introduces discrete token representations through vector quantization [36]. TopV [32], FastV [8], ToMe [5], HoliTom [23], and DART [28] further explore pruning or merging strategies to accelerate inference. These methods primarily optimize general efficiency, while our work emphasizes interaction centered evidence selection in egocentric streaming.

Video Understanding Benchmarks. Several benchmarks have been developed for long video understanding, include LongVideoBench [29], LVBench [27], and MLVU [39], which evaluate long context video language reasoning. Online and streaming evaluation includes OVO-Bench [18], StreamingBench [17], IP-IBench [14], Inf-Streams-Eval from Streaming VLM [31], LiveSports-3K from LiveCC [7], and Eyes Wide Open which introduces ESTP Bench for proactive egocentric streaming QA [38]. Egocentric datasets provide rich interaction structure, including EPIC-KITCHENS-100 [9], Ego4D [10], Ego4D Goal-Step for hierarchical procedural understanding [25], Ego-Exo4D for ego and exo perspective skill understanding [11] and EgoProx [13] for egocentric interaction reasoning from a spatial perspective.

Despite this progress, most existing benchmarks emphasize final accuracy under offline access or simplified streaming conditions. Our benchmark builds on these datasets while enforcing streaming constraints in egocentric settings and explicitly evaluating confidence behavior over time. A detailed benchmark comparison is provided in the supplementary appendix.

3 Problem Formulation and Benchmark Design

In what follows, we first present a temporally structured formulation of streaming interaction understanding. We then formalize the conditions under which a response is warranted under partial observability. Finally, we describe our evaluation tasks and their metrics, and present our benchmark construction.

3.1 Problem Formulation

Streaming interaction understanding requires MLLMs to produce timely and accurate judgments from continuously arriving video frames. Unlike conventional offline video understanding under full observability, this setting introduces response-timing challenges under partial observability. Specifically, models must determine (1) interaction presence — whether an interaction is occurring at all to warrant responding to the user's query, (2) the answerability of the situation — whether the inference is feasible given the currently available partial observation, and (3) the confidence level of the response under streaming uncertainty. Importantly, these challenges are inherently tied to distinct temporal formulations of the reasoning tasks.

Formally, we denote a streaming video as a sequence of short video clips $X_{1:t} = \{x_1, \ldots, x_t\}$, where x_t denotes the clip at time step t and $X_{1:t}$ are thus the observed clips up to t. Under a standard multiple-choice question (MCQ) answering setting, given a query Q (often in text form) and a candidate set C, the MLLM f_θ produces a response $\hat{A} = \mathcal{F}_\theta(Q, X_{1:t}) \in C$. We consider the three key reasoning tasks in streaming interaction understanding.

Present modeling. The query Q pertains to the current clip x_t, and the model is required to address Q using only x_t, without accessing to the past clips $X_{1:t-1}$ preceding x_t. In this setting, the model must explicitly assess the presence of interaction related to Q in the current clip x_t.

Prospective modeling. The query Q concerns a future clip $x_{t+\tau}$ with $\tau > 0$, and the model is required to address Q using only the observed clips $X_{1:t}$, without accessing to future clips $X_{t+1:t+\tau}$. Notably, the model must explicitly reason about if reliable prediction is feasible given the current observations.

Retrospective modeling. The query Q refers to a past clip $x_{t-\tau} \in X_{1:t}$, and the model must resolve Q based on the accumulated observations $X_{1:t}$. Naturally, the model is required to identify and retrieve the relevant prior context associated with $x_{t-\tau}$ from the observed history.

Note that we adopt the terms *prospective* and *retrospective* modeling to emphasize the streaming setting, where responses must be generated under partial observability and appropriate response timing must be determined. This distinguishes our formulation from conventional future anticipation or memory retrieval tasks [10].

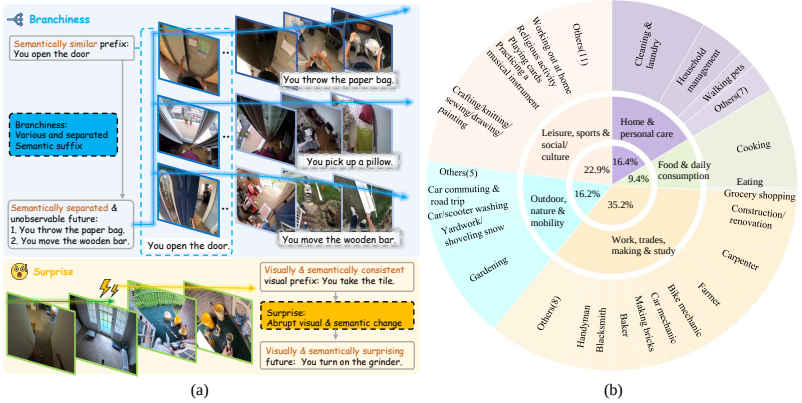

Fig. 2: Scenario distribution of EgoSAT and examples of predictability.(a) We attribute unpredictability to *branchiness* (a fixed, observable semantic prefix followedd by various and separated semantic suffix) and *surprise* (abrupt visual and semantic change). (b) Our EgoSAT covers 56 different scenarios categorized into five distinct activity groups.

3.2 Answerability and Confidence in Prospective Modeling

Answerability. A key to scientific evaluation in prospective modeling is to distinguish models' capabilities from inherent unpredictability, which derives from partial observation of the activities. Here we assume that the interaction categories and intervals are known, and we carefully characterize the answerability of future events via surprise and branchiness, as shown in Figure 2 (a). We further validate the answerability labels with a human sanity check in the supplementary material, where human judgments agree with our branchiness and surprise labels by 78% and 84%, respectively.

(1) Surprise: *unpredictability induced by abrupt local visual and semantic shifts.* We quantify *surprising* short-horizon futures by measuring the **cross-modal discrepancy level** between recent context and the imminent future. Specifically, for each query time t, we define a context window $A = [t - \tau, t)$ and a target window with the ground-truth event after t: $B = [t, t + h]$. We compute two complementary shift signals. **Visual shift:** we extract CLIP image features for frames within A and B, average-pool them followed by ℓ_2-normalize to obtain v_A and v_B, and compute the visual similarity $s_v = \cos(v_A, v_B)$. **Semantic shift:** we encode action texts with a CLIP text encoder; for all actions $\{a_i\}$ overlapping A, we assign overlap-based weights w_i with $\sum_i w_i = 1$, and compare them to the ground-truth text vector t_B via a weighted similarity $s_t = \sum_i w_i \cos(t_{a_i}, t_B)$. Because s_v and s_t come from different modalities and are not directly comparable in scale, we align them through an empirical distribution transform (probability integral transform + probit). Let $\hat{F}_v$ and $\hat{F}_t$ denote the empirical CDFs estimated over all samples; we map

$$p_v = \hat{F}_v(s_v), \quad p_t = \hat{F}_t(s_t), \qquad z_v = \Phi^{-1}(p_v), \quad z_t = \Phi^{-1}(p_t), \tag{1}$$

where Φ^{-1} is the inverse CDF of the standard normal (probit). We then fuse the two modalities with equal weights,

$$z = (z_v + z_t)/2, \qquad \text{Surprise}(t) = -z, \tag{2}$$

where the negative sign ensures that **lower similarity (stronger mismatch)** corresponds to **higher surprise** (low percentile $\Rightarrow$ negative z). This definition provides a parameter-free, reproducible cross-modal surprise score, enabling answerability-aware stratification and subsequent confidence-based diagnostics.

(2) Branchiness: *unpredictability induced by diverse and semantically dispersed successor interactions.* For each query time t, let O denote the ongoing interaction covering t, and let A be the earliest successor interaction whose start time falls within a short window $(t, t+h]$. Aggregating such transitions over the dataset yields a conditional successor distribution $p(A \mid O)$. We characterize this distribution with two complementary components: (i) **variety and evenness** via the entropy

$$S(O) = -\Sigma_i \; p_i \log p_i, \quad p_i = p(A_i \mid O), \tag{3}$$

and (ii) **semantic dispersion** via Rao's quadratic entropy

$$Q(O) = \Sigma_i \Sigma_j \; p_i p_j \, d(A_i, A_j), \tag{4}$$

where $d(A_i, A_j) = 1 - \cos(e_i, e_j)$ is the CLIP-text distance between successor interaction texts with their ℓ_2-normalized embeddings e_i. We then define the branchiness score

$$\text{Branchiness}(O) = B(O) = S(O) \cdot Q(O). \tag{5}$$

Intuitively, branchiness is high when the future has both many plausible successors (high S) and these successors are semantically far apart (high Q), indicating low predictability. This measure is complementary to Surprise: while Surprise captures abrupt evidence shifts, Branchiness captures multi-modal, multi-branch continuations, and together they enable answerability-aware evaluation and confidence diagnostics.

Confidence. Unpredictability induced by **branchiness** and **surprise**, together with evidence insufficiency in prospective modeling and evidence fade in retrospective modeling, contributes to streaming uncertainty. We therefore introduce *confidence diagnostics* to evaluate whether a model can correctly judge the feasibility of an inference under partial observability.

Following previous work on analyzing multiple-choice confidence scoring for LLMs [26], we quantify MCQ confidence by probing the model's next-token distribution over the forced one-letter selection set C. Let $p_{\text{raw}}(k)$ be the probability mass assigned to letter k (aggregating its single-token variants), $m = \sum_{k \in C} p_{\text{raw}}(k)$, and $p(k) = p_{\text{raw}}(k)/m$ be the conditional distribution on C. We define confidence and uncertainty as:

$$\text{Conf} = \max_{k \in C} \; p(k), \qquad \text{Ent} = -\Sigma_{k \in C} \; p(k) \log p(k), \tag{6}$$

where a low Conf (or high Ent) indicates low feasibility and enables predictability-/retrievability-aware confidence diagnostics in streaming evaluation.

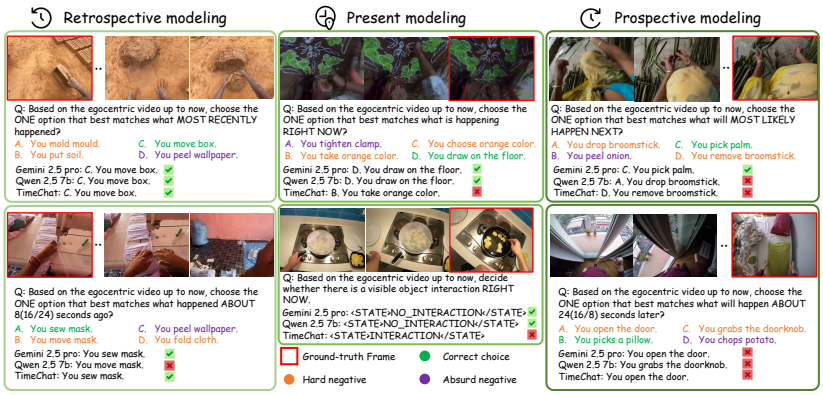

Fig. 3: Samples from our six task settings. Red border: the ground-truth frame. Choices are colored accordingly. Notably, the MCQ choices are *shuffled* to ensure minimal information leakage. We also show the model outputs.

3.3 Task Formulation and Metric Design

We now present the evaluation tasks under our problem formulation. For most tasks, we use multiple-choice accuracy as the primary metric to ensure objective and deterministic evaluation. We additionally report open-ended evaluation using LLM-based judging in the supplement. Implementation of these metrics is described in Sec. 4.2.

Present modeling tasks. For egocentric streaming assistants, real-time understanding via *present modeling* is essential not only for stable online interaction, but also as the perceptual foundation for retro-/prospective reasoning and memory. We focus on two tasks to evaluate this capability:

1. **Now narration**: Recognize the visibility of human-object interaction and what the interaction is.
2. **State switch**: Promptly adjust output state when the interaction state switches (interaction to no-interaction, or the reversal).

Prospective modeling tasks. Under partial visibility and uncertainty, *prospective modeling* presents significant challenges to streaming assistants. We introduce two tasks that evaluate the capability of utilizing partial observation for future prediction.

1. **Short-horizon anticipation**: Given streaming clip at time step t, reason over the observed visual prefix to predict the interaction that will occur in the near future at $t + \tau$.
2. **Multi-step anticipation:** Assuming an action begins at $t + \tau$, we probe whether the model can anticipate this event from progressively earlier visual prefixes ending at t, $t - \tau$, and $t - 2\tau$.

Retrospective modeling tasks. Retrospective modeling based on observed history, is another vital demonstration of streaming interaction understanding. We consider the following tasks for evaluating this capability.

1. **Short-horizon retrieval**: Retrieve information about the past clip $x_{t-\tau}$ from the streaming inputs
2. **Multi-step retrieval**: Retrieve an anchored past event at several past steps forming an arithmetic sequence: $t-\tau$, $t-2\tau$, and $t-3\tau$.

3.4 Benchmark Construction

Video source. Our benchmark is built on the recent effort in creating large-scale egocentric video datasets. We adopt Ego4D [10] as the video source. Ego4D is a massive-scale egocentric video dataset with over 3,600 hours of video collected around the world, covering a wide range of (>100) real-world activities and providing human annotated intervals for many hand-object interactions.

Data curation. In an egocentric setting, we define a segment as an interaction segment only when there is a visible hand-object interaction in the view; otherwise, we treat it as a *gap*, *i.e.*, a background segment between interactions. We leverage Ego4D's mature rejection tags as a quality-control mechanism and filter out events that do not contain visible interactions. Specifically, we use two rejection reasons: (1) *no human-object interaction from camera wearer*, and (2) *object of change is not visible*. We validate the reliability of these rejection tags in the appendix through manual inspection. Finally, we orchestrate 1,997 video sessions with valid annotations, spanning 56 scenarios. Each session lasts roughly 230–300 seconds, with an average duration of 296.8 seconds. These sessions typically contain dense hand-object interactions, making them adequate for constructing streaming QA with interaction-centric ground truth.

QA pair construction. Since our QA ground truth is derived from Ego4D annotations and filtered using the above quality-assurance rules, we construct queries using fixed templates with dynamically generated answer options for MCQ evaluation. We additionally provide an open-ended version of all queries using the same templates, enabling complementary evaluation when needed.

Each MCQ instance contains four options: the ground-truth answer, two hard negatives, and one absurd negative. The hard negatives are designed to be temporally confusing and are sampled from events adjacent to the ground-truth segment while ensuring different *(verb, noun)* pairs. Specifically, for Now Narration task, negatives are drawn from the nearest neighboring events of the ground-truth event. For Short-horizon Anticipation and Retrieval tasks, we sample negatives around the *(query time, ground-truth segment)* pair. For Multi-step Anticipation and Retrospection, negatives are sampled near the query timestamp, as queries and answers may be temporally far apart.

The absurd negative is designed to be semantically distant from the ground truth. We construct a candidate pool of *(verb, noun)* pairs from the training split for each task, remove the top 10% most frequent pairs to avoid trivial options, and exclude pairs appearing in the same video as the ground truth. The absurd negative is then selected as the candidate with the largest text-embedding distance to the ground-truth option (see supplement for details).

Finally, we randomly shuffle the four options for each MCQ instance and store the shuffled queries offline, following prior benchmark practices to mitigate answer-position bias and potential information leakage. We also include a *blind baseline* using a text-only model to verify that the answer options alone provide little information beyond random guessing.

MCQ validity checks. To ensure that EgoSAT evaluates visual-temporal reasoning rather than artifacts of the multiple-choice format, we further conduct blind and distractor analyses in the supplementary material. Blind models that receive only the questions and shuffled options suffer clear performance drops on MCQ-based temporal reasoning tasks, indicating that the options alone do not reveal the answer. Moreover, when models make incorrect predictions, they predominantly select temporally plausible hard negatives rather than semantically absurd negatives, and this trend remains stable after re-shuffling answer positions. These results support that EgoSAT provides strong distractors while avoiding strong text-only or option-position shortcuts.

EgoSAT. Our final benchmark, EgoSAT, contains 1,997 unique videos spanning 165 hours of egocentric footage and around 4,800 high-quality question–answer pairs. The distribution of videos across scenarios are shown in Figure 2 (b). In terms of task-level distribution, our benchmark is evenly distributed by design.

4 Experiments and Results

4.1 Experiment Protocol

Models. We evaluate three groups of models (Table 1).

1. *Offline video LLMs*, including two proprietary models (Gemini 2.5 Pro [2], Claude Sonnet 4 [1]) and several open-weight models (Qwen2.5-VL [4], Video-LLaVA [16]).
2. *Streaming VLM*, TimeChat-Online-7B [35], which performs streaming inference via dynamic KV cache dropping, Flash-VStream [37], and LLaVA-OV1.5 [3].
3. *ROI-augmented Streaming VLM*, our variant (ROI+TimeChat-Online) of TimeChat-Online that uses egocentric hand and gaze cues for token selection.

Offline-online emulation. Since most off-the-shelf video LLMs only support offline processing, and do not accept streaming inputs, we simulate strict-online evaluation by truncating the video and retain only the visual prefix at each query time t. This ensures no look-ahead for all offline baselines.

Our training-free ROI module. One of the models evaluated in our study, TimeChat-Online, introduces Dynamic Token Drop (DTD) to improve efficiency in streaming settings by mitigating temporal and spatial redundancy in visual tokens. We observe that, in egocentric streaming inference, informative visual regions often concentrate around hand–object interactions and gaze-related areas. Motivated by this observation, we define an Interaction Region of Interest (ROI) that captures the spatial neighborhoods surrounding these egocentric

cues. Building on this idea, we introduce ROI-TimeChat-Online, an ROI-aware variant of TimeChat-Online that prioritizes KV cache of tokens within the Interaction ROI during token selection. Importantly, this modification is parameter-free and training-free, making it easy to incorporate into existing models. We evaluate ROI-TimeChat-Online on our benchmark to study how egocentric priors can improve streaming interaction understanding.

Supervised fine-tuning (SFT). Beyond zero-shot evaluation, we apply lightweight LoRA-based SFT to the two TimeChat-Online variants using next-token LM loss on EgoSAT target outputs. This setting is used to reduce schema violations and examine whether the observed ROI-related trends persist after task-format alignment.

Metrics. We instantiate the metrics designed in Sec. 3.3. Further details can be found in the supplement. *Now Narration Task*: we also report the recall for interaction-visible class. *State Switch Task*: we anchor the query time before switch and gradually moving the After-switch query to the GT Switch moment. Here, a successful state switch occurs if and only if the model predicts correct states in both the two queries. We then report the rate of success when the After-switch query is furthest to the GT switch moment. *Short-horizon Anticipation Task*: we compare confidence scores from the models against those derived in Sec. 3.2. *Multi-step Anticipation Task*: we divide queries into three groups based on their relative distance to anchor events: lead $\tau(2\tau, 3\tau)$ means GT event is $\tau(2\tau, 3\tau)$ seconds ahead ($\tau = 8$), and break down the MCQ accuracy. For *Short-horizon / Multi-step Retrieval Tasks*, we report metrics similar to Prospective Modeling, except that we do not associate answerability with Short-horizon Retrospection in our task settings.

4.2 Results on EgoSAT Benchmark

Table 1 reports the main results across present, prospective, and retrospective modeling tasks. Overall, proprietary frontier models such as Gemini 2.5 Pro and Claude Sonnet 4 achieve the strongest performance on present narration tasks, suggesting that large-scale pretraining and proprietary data pipelines still provide general advantages even in the challenging egocentric scenarios considered in our benchmark. Interestingly, proprietary models do not exhibit clear advantages on the precision metric for narration generation. We attribute this behavior to human preference alignment objectives, which encourage models to produce reasonable responses even in partially unanswerable settings, leading to lower precision under the strict evaluation protocol adopted in our benchmark.

In addition, both prospective and retrospective modeling pose significant challenges for prevailing methods, albeit in different ways. Retrospective reasoning requires models to retrieve relevant information at the correct temporal point, which can be hindered by imperfect temporal grounding, information loss during sampling, and positional bias introduced by temporal embeddings. On the other hand, prospective modeling introduces inherent uncertainty, as the future outcome often represents only one of many plausible possibilities. This observa-

Table 1: Main results (MCQ-first). All models are evaluated with strict-online prefix truncation at each query time t. **Rec.** is recall for interaction-visible (TP/(TP + FN)), and **Prec.** is precision (TP/(TP + FP)) **Multi.** reports average MCQ accuracy over three probes at $\tau, 2\tau, 3\tau$.

Model\task	Present					Prospective		Retrospective	
	Narration			State switch		Short	Multi.	Short	Multi.
	Rec.	Prec.	MCQ acc	Fg→bg	Bg→fg	MCQ acc.	MCQ avg. acc.	MCQ acc.	MCQ avg. acc.
Human agents									
Human	84.62	93.13	73.13	60.63	63.75	70.63	63.13	76.25	73.75
Offline proprietary models									
Gemini 2.5 Pro	69.63	89.02	43.00	21.43	16.67	29.92	29.43	39.80	48.15
Claude Sonnet 4	77.51	86.30	38.43	11.59	12.00	32.20	29.98	40.46	39.64
Offline open-sourced models									
Qwen2.5-VL-72B	73.66	85.64	38.86	3.95	8.64	43.06	26.00	51.87	43.46
Qwen2.5-VL-32B	86.16	86.94	39.91	2.63	7.41	37.24	25.23	50.86	40.74
Qwen2.5-VL-7B	78.72	84.78	36.50	4.00	6.25	27.84	28.11	42.79	34.46
Video-LLaVA 7B	100.00	83.46	25.75	0.00	0.00	25.11	25.50	20.89	25.00
Blind model: see Table 4: Blind Model									
Online models									
TimeChat-Online-7B	98.51	83.35	33.33	0.00	0.00	29.72	26.01	37.84	31.63
ROI-TimeChat-Online	97.92	83.40	36.47	0.00	1.25	30.51	25.05	34.41	30.82
Flash-VStream	94.61	87.44	36.92	1.02	1.78	29.64	28.50	39.76	33.80
LLaVA-OV1.5	93.27	84.68	32.30	2.54	2.03	25.53	23.68	37.84	32.39
SFT Online Models									
TimeChat-Online-7B	0.00	0.00	45.59	0.00	0.00	61.63	42.79	43.90	45.43
ROI-TimeChat-Online	0.00	0.00	46.94	0.00	0.00	59.54	40.55	44.00	41.24

tion further emphasizes the importance of explicitly accounting for predictability in anticipation settings.

Online vs. offline modeling. We further compare streaming online TimeChat-Online-7B models with offline models of similar scale. Despite operating at comparable model capacity, the online model consistently lags behind its offline counterpart across most tasks. This performance gap primarily stems from the streaming memory constraint: online models must maintain a compressed representation of past observations through incremental caching, which inevitably leads to information loss. As expected, the gap becomes more pronounced in multi-step settings, where longer temporal dependencies exacerbate the information loss accumulated during caching. A complementary memory-proxy study in the supplementary further shows that simply increasing sparse historical context does not yield monotonic gains, suggesting that effective streaming understanding requires adaptive memory selection rather than merely longer context windows.

ROI-based KV caching and task-specific SFT. For a fair comparison, we enforce the same KV caching budget for both TimeChat-Online and its ROI-

augmented version. Interestingly, this simple ROI caching improves model performance on present modeling tasks by 3.14%. However, it leads to large performance degradation on both prospective and retrospective modeling tasks. We attribute this to the fact that these tasks often require reasoning over broader contextual cues in the background to address temporal dynamics, whereas the hand- and gaze-conditioned ROI regions may omit such cues.

Unsurprisingly, task-specific SFT yields notable performance improvements across tasks, especially for prospective modeling. We freeze the visual encoder during SFT following standard practice. The substantial gains suggest that existing models are already capable of extracting temporally sensitive visual representations, but lack the instruction-level alignment needed to connect these representations with temporally grounded MCQ reasoning tasks. Lightweight SFT can effectively address this misalignment.

For SFT variants, some entries are *(near) zero* in Table 1 because the MCQ-first SFT training set only has MCQ type data with structured output schema, which can catastrophically violate the state-only output format required by the interaction-visibility evaluation in Now State Switch. We therefore report additional results for this metric in the supplementary. We additionally verify in the supplementary material that *SFT gains transfer across held-out scenario splits*, suggesting that the improvements are not merely due to in-scenario overfitting.

Human Baseline From above, we see huge performance gap in our task settings, especially Prospective Modeling and Retrospective Modeling, even for proprietary MLLMs. We therefore conduct comparative experiments on human agents to validate the feasibility or upper-bound of our tasks. We uniformly sample 10% of the multiple-choice questions from each of our 6 task to form a validation set for human agents.

As shown in Table 1, humans achieve strong performance across present, prospective and retrospective tasks, substantially surpassing all evaluated models. This gap indicates that EgoSAT is far from saturated and that low model performance primarily reflects capability limitations rather than inherent unanswerability of the benchmark. Meanwhile, human accuracy remains lower on state switch and multistep prediction, highlighting interaction-boundary sensitivity and observation insufficiency as key challenges in streaming interaction understanding.

4.3 Answerability Diagnostics

Confidence-level for state switch task. Interaction boundaries pose a key challenge in streaming settings, where models must switch their output state according to interaction visibility. As shown in Table 1, although proprietary models achieve relatively better state-switch accuracy, all tested models remain far from satisfactory. We provide a fine-grained analysis of query timing around the ground-truth switch moment in the supplementary material. The supplementary timing analysis further shows that models respond asymmetrically to

Table 2: Detailed results of prospective and retrospective tasks. We report confidence slope and accuracy by temporal distance. *Conf slope*: mean least-squares slope of confidence vs. temporal distance over the three probes; a negative slope implies confidence decreases with the temporal distance.

Model	Multi-step anticipation				Multi-step retrospection			
	Acc lead 3τ	Acc lead 2τ	Acc lead τ	Conf slope	Acc lag τ	Acc lag 2τ	Acc lag 3τ	Conf slope
Qwen-2.5-VL-72B	27.99	23.51	26.49	-0.76	47.78	42.22	40.37	-0.16
Qwen-2.5-VL-32B	25.68	22.07	27.93	<0.01	42.96	40.37	38.89	<0.01
Qwen-2.5-VL-7B	27.86	29.39	31.30	0.16	40.00	37.78	37.41	0.86
Video-LLaVA-7B	24.63	23.13	26.49	1.24	28.15	24.81	24.81	1.44
TimeChat-Online-7B	27.99	26.49	29.85	0.61	44.07	42.96	42.59	0.99
ROI-TimeChat-Online	29.09	20.19	25.88	17.36	30.05	28.30	34.10	4.61
TimeChat-Online-7B (SFT)	42.91	42.91	42.54	<0.01	45.93	48.52	41.85	0.01
ROI-TimeChat-Online (SFT)	38.06	42.16	41.42	-0.07	42.47	44.02	44.40	0.40

Table 3: Detailed results of short-horizon anticipation: accuracy and confidence statistics on predictable vs. unpredictable queries.

Model	Short-horizon anticipation							
	predictable				unpredictable			
	accuracy	conf	conf_correct	conf_wrong	accuracy	conf	conf_correct	conf_wrong
Qwen2.5VL-72B	37.56	68.43	72.07	66.24	28.11	73.00	74.02	72.61
Qwen2.5-VL-32B	38.05	65.99	68.50	64.45	31.07	66.68	69.87	65.24
Qwen2.5-VL-7B	34.31	53.01	54.90	52.03	24.26	51.26	47.58	52.44
Video-LLaVa-7B	18.37	52.24	53.12	52.04	19.23	53.91	53.89	53.92
TimeChat-Online-7B	33.50	49.61	51.46	48.68	28.99	48.81	49.76	48.42
ROI-TimeChat-Online	32.58	49.14	48.90	49.25	27.16	48.40	46.49	49.11
TimeChat-Online-7B (SFT)	57.24	52.16	58.53	43.62	47.04	48.82	52.97	45.13
ROI-TimeChat-Online (SFT)	47.10	44.34	49.75	39.52	35.46	42.28	47.57	39.38

state transitions, with entering interactions generally easier to detect than exiting them, highlighting the difficulty of maintaining calibrated interaction-state beliefs.

Confidence-level for multi-step tasks. For multi-step anticipation and retrieval, answerability is closely tied to temporal distance. Larger anticipation lead implies less relevant observation for predicting the anchored event, while larger retrieval lag makes past evidence more likely to be diluted or overwritten by later content. We therefore expect both accuracy and self-contained confidence to decrease as lead or lag grows. To test this, we report accuracy at three temporal distances and summarize the confidence–distance relationship by fitting a line to each model's three-point confidence curve, with the normalized mean slope reported in Table 2.

Across tasks, multi-step retrospection is generally easier than multi-step anticipation, since the queried events have already occurred and only need to be localized in the observed history. However, many models still fail to show monotonic accuracy degradation as lead/lag increases, especially for anticipation,

where local visual similarity and inherent uncertainty can dominate temporal distance. Moreover, confidence slopes are often inconsistent with the expected negative trend, indicating that current models' self-contained confidence does not reliably track factual answerability or temporal difficulty.

Predictability. As discussed earlier, predictability for short-horizon anticipation is shaped by the *branchiness* and *surprise* of interaction dynamics. For an uncertainty-aware AI agent, it should assign lower confidence to unpredictable queries than to predictable ones. Note that we evaluate predictability only on short-horizon tasks, as the multi-step setting is already sufficiently challenging; therefore, we restrict those tasks to predictable scenarios. We present the breakdown of predictable and unpredictable queries in Table 3. Not surprisingly, models generally achieve higher accuracy on predictable queries than on unpredictable ones. However, we observe little difference in the confidence levels assigned to these two categories. These results suggest that current MLLMs struggle to distinguish predictable from inherently uncertain situations, often producing overconfident responses even when predictions are incorrect, indicating that their self-contained confidence does not reliably reflect the factual predictability of the query under high unanswerability.

5 Conclusion

This paper introduced **EgoSAT**, the first comprehensive benchmark for egocentric streaming interaction understanding. We formulated streaming egocentric video reasoning through a unified temporal perspective, evaluating vision–language models across three regimes corresponding to **retrospective**, **present**, and **prospective** events. Beyond temporal reasoning, our benchmark also examines a fundamental property of streaming interaction—*answerability*—by evaluating both the confidence calibration of model responses and the inherent predictability of future events.

Our results reveal consistent performance gaps between offline and streaming settings, highlighting the challenges current models face in maintaining temporally grounded representations under constrained KV-cache budgets. We believe that **EgoSAT** provides a principled testbed for studying streaming video understanding and offers new insights into the temporal reasoning limitations of modern vision–language models. More broadly, our benchmark and analysis lay the groundwork for developing proactive AI assistants capable of continuous perception and reasoning in real-world egocentric environments.

Acknowledgments

This work is supported by Tsinghua University–Keystone Electrical (Zhejiang) Co., Ltd. Joint Research Center for Embodied Multimodal Artificial Intelligence (JCEMAI).

References

1. Claude opus 4 and claude sonnet 4 summary table (2025), https://www.anthropic.com/transparency
2. Gemini 2.5 pro model card (2025), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf
3. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Didi, Z., Wu, C., Tan, H., Li, C., Yang, J., Yu, J., Wang, X., Qin, B., Wang, Y., Yan, Z., Feng, Z., Liu, Z., Li, B., Deng, J.: Llava-onevision-1.5: Fully open framework for democratized multimodal training (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
5. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: The Eleventh International Conference on Learning Representations
6. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18407–18418 (2024)
7. Chen, J., Zeng, Z., Lin, Y., Li, W., Ma, Z., Shou, M.Z.: Livecc: Learning video llm with streaming speech transcription at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29083–29095 (2025)
8. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35 (2024)
9. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision **130**(1), 33–55 (2022)
10. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
11. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024)
12. Huang, Y., Xu, J., Pei, B., He, Y., Chen, G., Yang, L., Chen, X., Wang, Y., Nie, Z., Liu, J., et al.: Vinci: A real-time embodied smart assistant based on egocentric vision-language model. arXiv preprint arXiv:2412.21080 (2024)
13. Li, J., Chen, Y., Piao, D., Pan, P., Yu, Y., Wang, D., Yan, H., Yue, L., Wang, S., Chen, Y., et al.: Egoprox: Evaluating mllms on egocentric 3d proximity reasoning across a cognitive hierarchy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23751–23762 (2026)
14. Li, J., Chen, Y., Song, W., Lei, Y., Zhang, Y., Yan, H., Pan, P., Liu, M.: Ipibench: Evaluating interactive proactive intelligence of mllms under continuous streams. arXiv preprint arXiv:2605.27074 (2026)

15. Li, X., Wang, Y., Yu, J., Zeng, X., Zhu, Y., Huang, H., Gao, J., Li, K., He, Y., Wang, C., et al.: Videochat-flash: Hierarchical compression for long-context video modeling. arXiv e-prints pp. arXiv–2501 (2024)
16. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2024), https://arxiv.org/abs/2311.10122
17. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024)
18. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18902–18913 (2025)
19. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24045–24055 (2025)
20. Qian, R., Dong, X., Zhang, P., Zang, Y., Ding, S., Lin, D., Wang, J.: Streaming long video understanding with large language models. Advances in Neural Information Processing Systems **37**, 119336–119360 (2024)
21. Ren, S., Chen, S., Li, S., Sun, X., Hou, L.: Testa: Temporal-spatial token aggregation for long-form video-language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 932–947 (2023)
22. Ryoo, M.S., Piergiovanni, A., Arnab, A., Dehghani, M., Angelova, A.: Tokenlearner: What can 8 learned tokens do for images and videos? arXiv preprint arXiv:2106.11297 (2021)
23. Shao, K., Keda, T., Qin, C., You, H., Sui, Y., Wang, H.: Holitom: Holistic token merging for fast video large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
24. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: Longvu: Spatiotemporal adaptive compression for long video-language understanding. In: 42nd International Conference on Machine Learning, ICML 2025. ML Research Press (2025)
25. Song, Y., Byrne, E., Nagarajan, T., Wang, H., Martin, M., Torresani, L.: Ego4d goal-step: Toward hierarchical understanding of procedural activities. Advances in neural information processing systems **36**, 38863–38886 (2023)
26. Tsvilodub, P., Wang, H., Grosch, S., Franke, M.: Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods. arXiv preprint arXiv:2403.00998 (2024)
27. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22958–22967 (2025)
28. Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., Zhang, L.: Stop looking for "important tokens" in multimodal language models: Duplication matters more. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9972–9991 (2025)
29. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)

30. Xiong, H., Yang, Z., Yu, J., Zhuge, Y., Zhang, L., Zhu, J., Lu, H.: Streaming video understanding and multi-round interaction with memory-enhanced knowledge. In: The Thirteenth International Conference on Learning Representations
31. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025)
32. Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Li, C., Yan, J., Bai, Y., Sadayappan, P., Hu, X., et al.: Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19803–19813 (2025)
33. Yang, C., Dong, X., Zhu, X., Su, W., Wang, J., Tian, H., Chen, Z., Wang, W., Lu, L., Dai, J.: Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24939–24949 (2025)
34. Yao, L., Li, L., Ren, S., Wang, L., Liu, Y., Sun, X., Hou, L.: Deco: Decoupling token compression from semantic abstraction in multimodal large language models. arXiv preprint arXiv:2405.20985 (2024)
35. Yao, L., Li, Y., Wei, Y., Li, L., Ren, S., Liu, Y., Ouyang, K., Wang, L., Li, S., Li, S., et al.: Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10807–10816 (2025)
36. Zhang, H., Fu, Y.: Vqtoken: Neural discrete token representation learning for extreme token reduction in video large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
37. Zhang, H., Wang, Y., Tang, Y., Liu, Y., Feng, J., Jin, X.: Flash-vstream: Efficient real-time understanding for long video streams. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21059–21069 (2025)
38. Zhang, Y., Shi, C., Wang, Y., Yang, S.: Eyes wide open: Ego proactive video-llm for streaming video. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
39. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13691–13701 (2025)