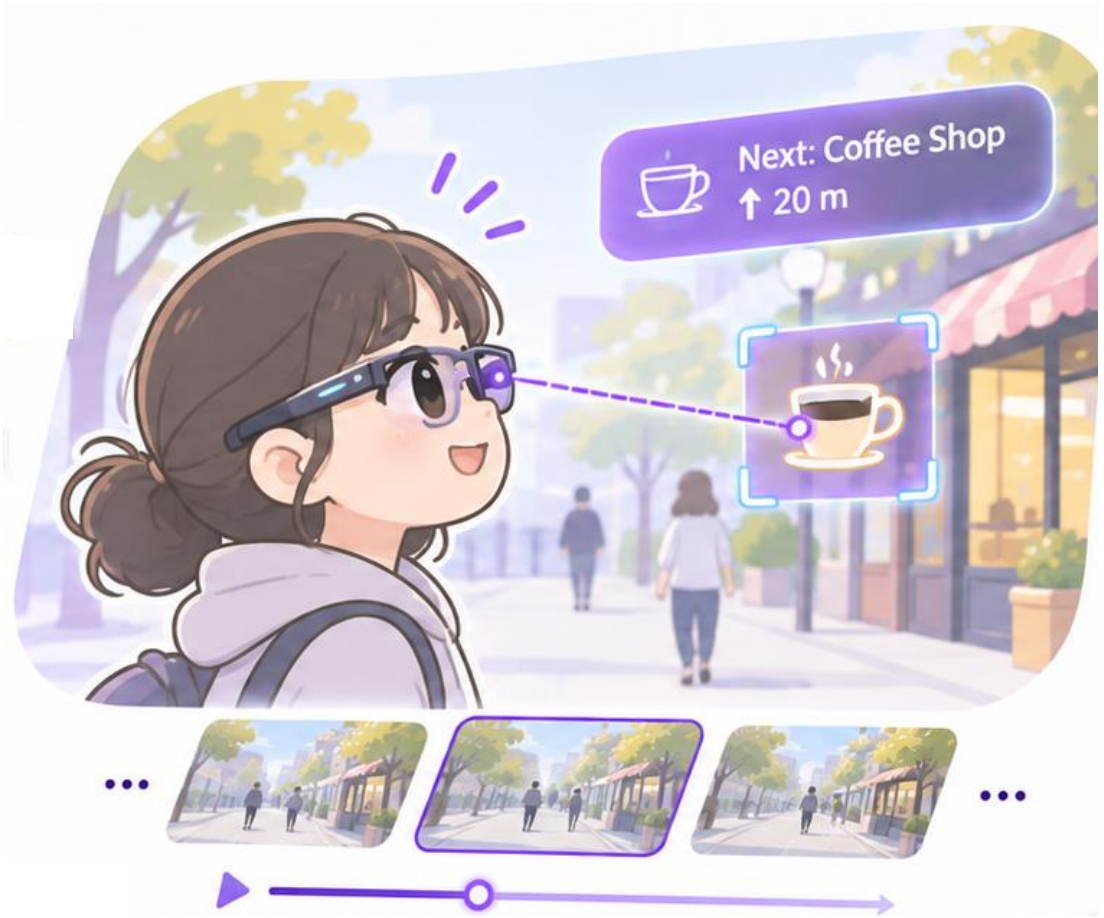
NeurIPS 2026

StreamGaze

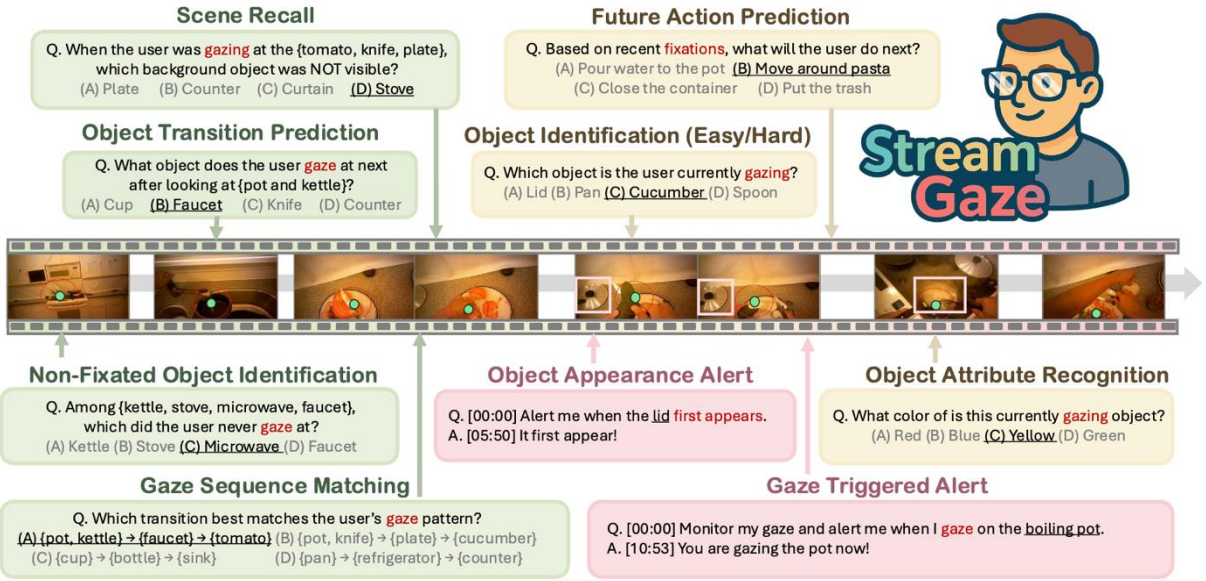
Gaze-Guided Temporal Reasoning and Proactive Understanding in Streaming Videos



Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton,
Ryan A. Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui,
Franck Dernoncourt, Mohit Bansal

Streaming Video Understanding with Gaze

- **Motivation:** While prior streaming benchmarks evaluate temporal reasoning, none measure whether MLLMs can interpret or leverage human gaze signals within a streaming setting.
- **Research Question:** We rely on gaze to guide our actions, but can current MLLMs truly understand it and infer our intentions?
- **Our proposal:** We introduce **StreamGaze**, the first benchmark/data generation pipeline designed to evaluate how effectively MLLMs use gaze for temporal and proactive reasoning in streaming videos.





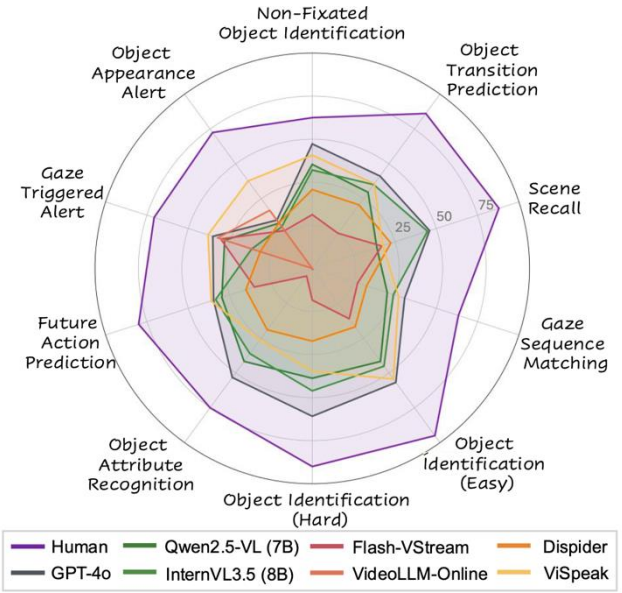
Our Contribution

1. **Gaze-Guided Streaming Benchmark**: 10 tasks spanning past, present, and proactive reasoning, from gaze-sequence matching to alerting when objects appear within the FOV area.

2. **Gaze-Guided Streaming Data Construction Pipeline**: We align egocentric videos with raw gaze trajectories using fixation extraction, region-specific visual prompting, and scanpath construction to generate spatio-temporally grounded QA pairs.

3. **Comprehensive Evaluation of State-of-the-Art MLLMs**: Across all gaze-conditioned streaming tasks, we highlight fundamental limits of current MLLMs.

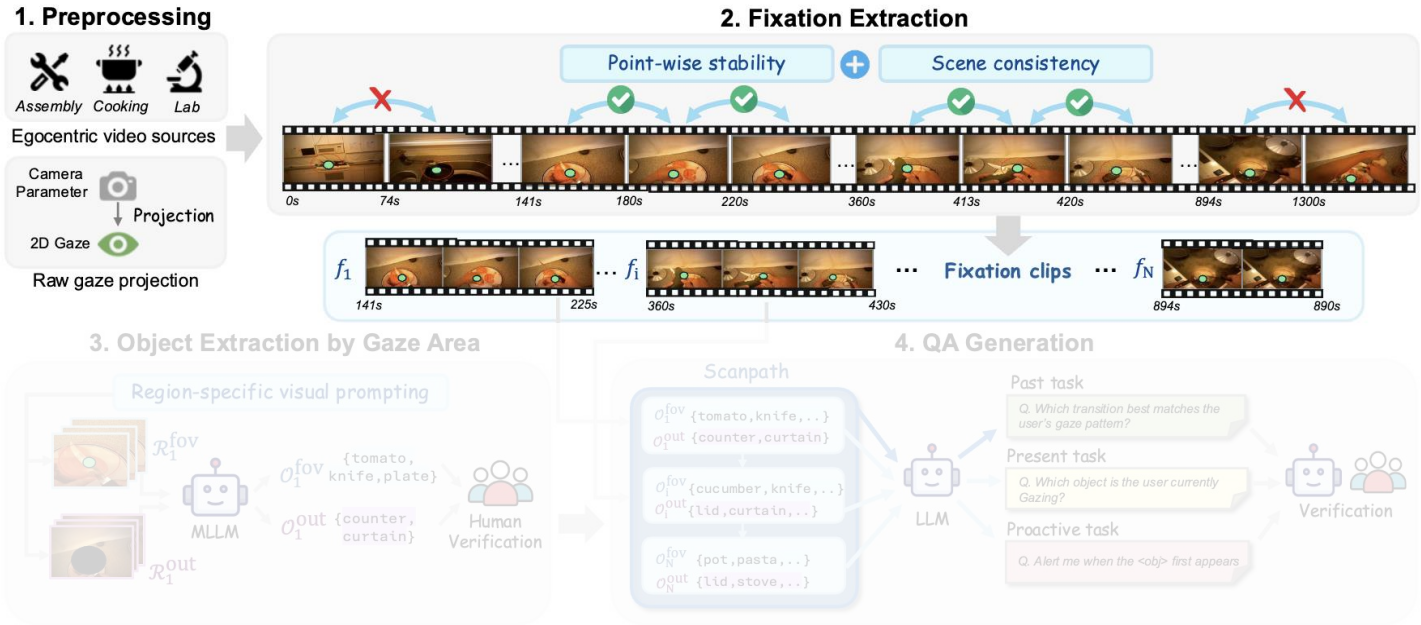
Dataset	Avg video (s)	QA pairs	QA type	Annotation	Proac.	Gaze	All-time	Ego.
Gaze-based QA benchmarks								
GazeVQA [7]	150	25040	MC	Human	✘	✓	✘	✓
EgoGazeVQA [19]	52	1757	MC	Auto	✘	✓	✘	✓
Streaming QA benchmarks								
StreamingBench [13]	268	4500	MC+OE	Auto & Human	✓	✘	✓	✘
SVBench [31]	147	7374	OE	Auto & Human	✘	✘	✘	△
OVO-Bench [15]	428	2814	MC+OE	Auto & Human	✓	✘	✓	△
OmniMMI [28]	324	2290	OE	Auto & Human	✓	✘	✘	△
ProAssist [36]	710	1020	OE	Auto	✓	✘	✘	✓
StreamBench [29]	270	1800	OE	Auto & Human	✘	✘	✘	✓
Vispeak-Bench [4]	21	1000	OE	Auto & Human	✓	✘	✘	✘
StreamGaze (Ours)	815	8521	MC+OE	Auto & Human	✓	✓	✓	✓



1. Gaze-guided Streaming Data Construction

Our gaze-video QA generation pipeline aligns egocentric videos with raw gaze trajectories through four stages:

- **(1) Input:** egocentric videos, 2D gaze axis per each frame (projected via camera parameters)
- **(2) Fixation Extraction:** We extract fixation moments, intervals where the gaze remains relatively stable within a localized region with two criteria: *(i) point-wise stability* and *(ii) scene consistency*



Algorithm 1 Gaze-Guided Data Construction Pipeline

Input: $\mathcal{V} = \{v_t\}_{1:T}$, $\mathcal{G} = \{(x_t, y_t)\}_{t=1}^{T}$, $r_{\text{thresh}}, \tau_{\text{dur}}, \tau_{\text{scene}}$

```
# Fixation Extraction
F ← ∅
for each candidate interval [t_i^s, t_i^e] do
    Compute centroid (x̄_i, ȳ_i)                                   Point-wise stability
    if ∀ t ∈ [t_i^s, t_i^e], ‖(x_t, y_t) − (x̄_i, ȳ_i)‖_2 ≤ r_thresh then
        if t_i^e − t_i^s ≥ τ_dur then
            S_min ← min_{t=t_i^s}^{t_i^e − 1} ρ(H_t, H_{t+1})      Scene consistency
            if S_min ≥ τ_scene then
                F ← F ∪ {(x̄_i, ȳ_i, t_i^s, t_i^e)}
            end if
        end if
    end if
end for
```

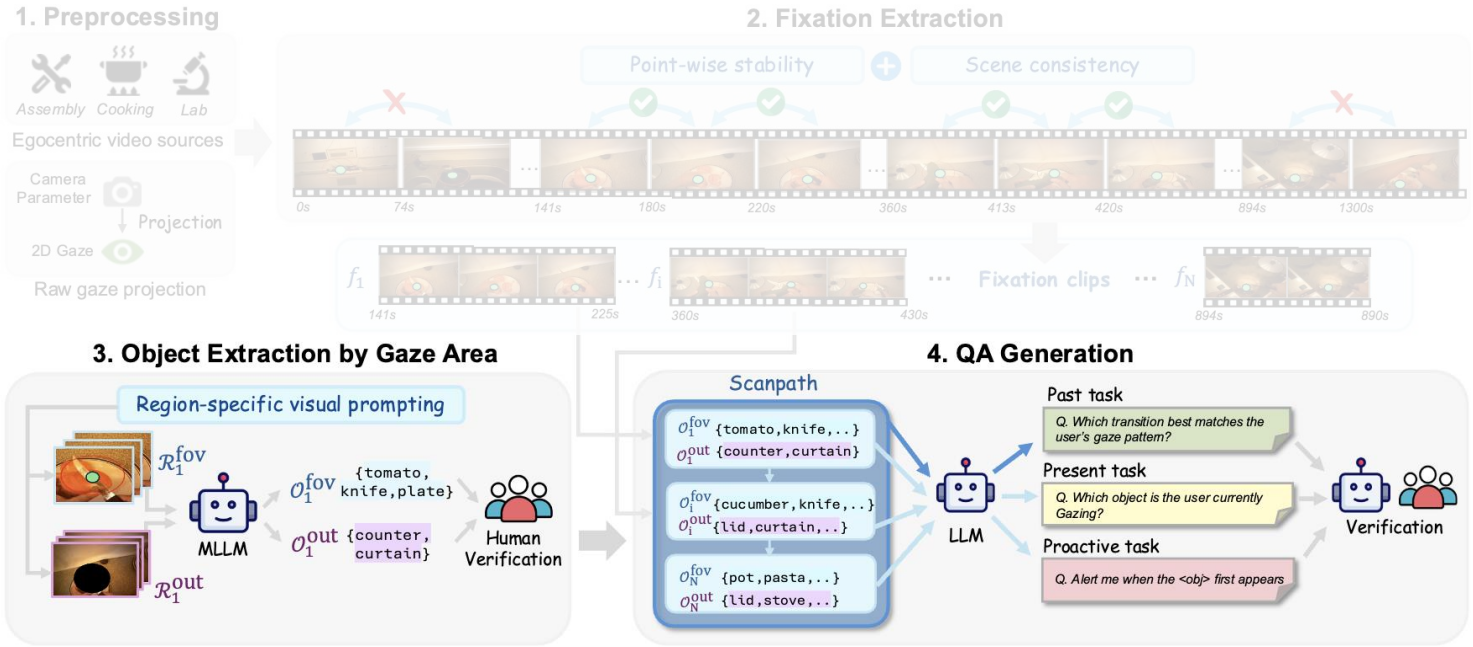
1. Gaze-guided Streaming Data Construction

Our gaze-video QA generation pipeline aligns egocentric videos with raw gaze trajectories through four stages:

- **(3) Region-Specific Visual Prompting:** We divide each frame into FOV and out-of-FOV regions and extract objects within the gaze area with MLLM.

$$\mathcal{R}_{i,t}^{\text{fov}} = \{(x, y) \mid \|(x, y) - (\bar{x}_i, \bar{y}_i)\|_2 \leq \tau_{fov}\},$$

- **(4) Scanpath Construction & QA Generation:** We construct a scanpath that represents the temporal evolution of gaze-guided object observations and generate QA pairs



Algorithm 1 Gaze-Guided Data Construction Pipeline

⋮

```
# Object and scanpath extraction
for each f_i ∈ F do
    for t = t_i^s ... t_i^e do
        Define R_{i,t}^fov and R_{i,t}^out
    end for
    O_i^fov ← MLLM(R_i^fov)
    O_i^out ← MLLM(R_i^out)
end for
S ← {(O_i^fov, O_i^out)}_{i=1}^N
# QA Generation
```

Object extraction

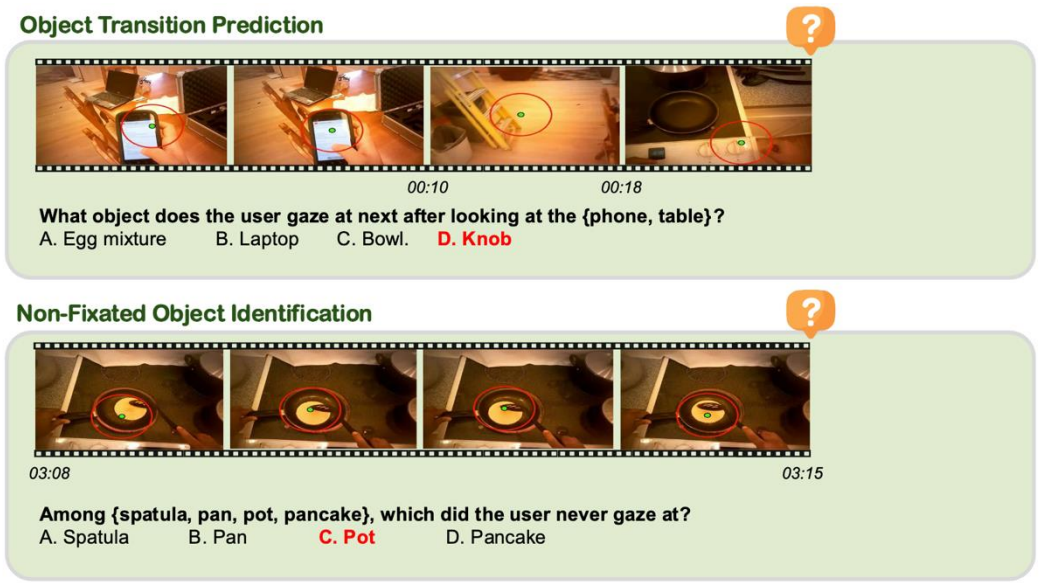
Scanpath construction

2. Benchmark Overview

$$\textbf{Past:} \quad A = \mathcal{M}\big(Q;\, \mathcal{V}_{[0,\,t_q]},\, \mathcal{G}_{[0,\,t_q]}\big)\,,$$
$$\textbf{Present:} \quad A = \mathcal{M}\big(Q;\, \mathcal{V}_{(t_q-\omega,\,t_q]},\, \mathcal{G}_{(t_q-\omega,\,t_q]}\big)\,,$$
$$\textbf{Proactive:} \quad A = \mathcal{M}\big(Q;\, \mathcal{V}_{(t_q,\,\infty]},\, \mathcal{G}_{(t_q,\,\infty]}\big)\,.$$

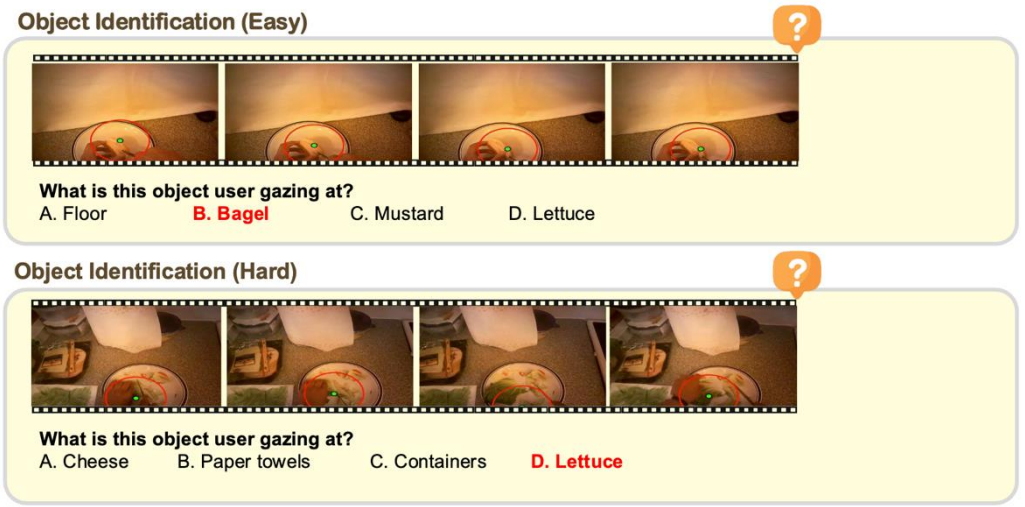
Problem setup: We formulate a gaze-guided streaming video understanding task, where the model must answer time- sensitive questions while observing temporally incoming frames and user gaze.

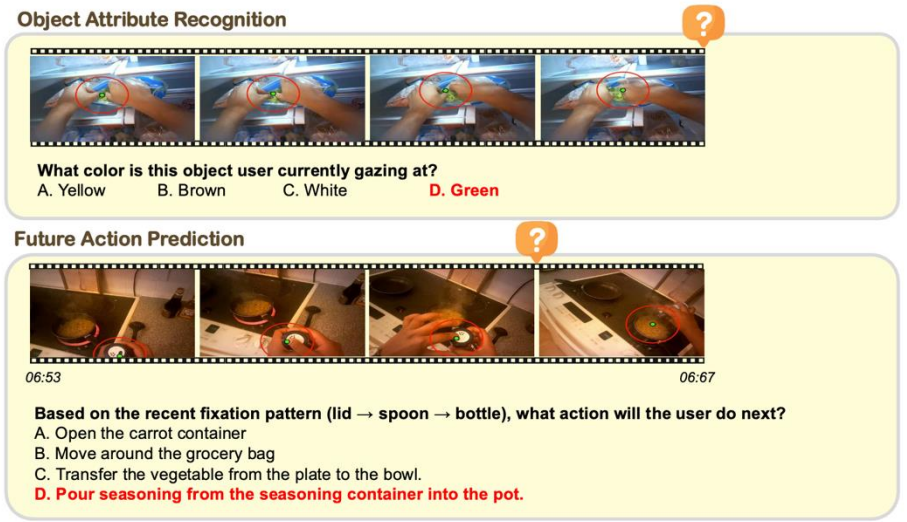
- **Past task:** focus on temporal reasoning over gaze and capture the dynamic characteristics of the user's gaze, modeling how attention shifts across objects over time.



2. Benchmark Overview

- **Present task**: capture the user's perceptual state and intention. We control question difficulty using FOV/out-of-FOV objects





- **Proactive task**: involve gaze-based reasoning, where the model anticipates or assists future user behavior based on gaze and contextual cues.

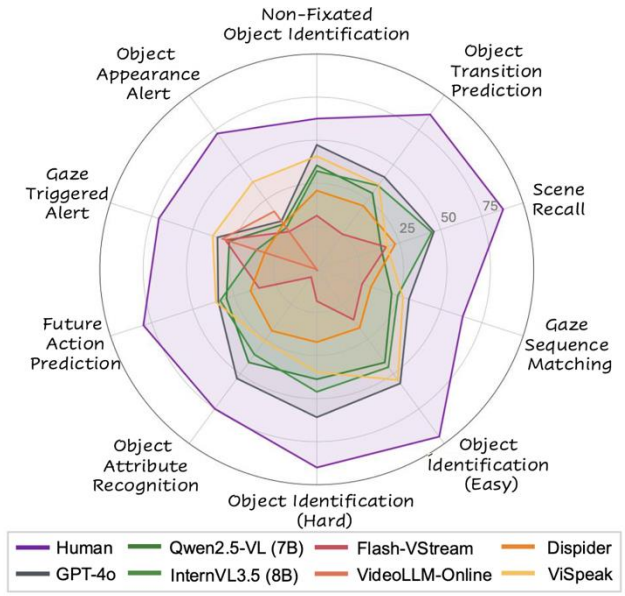




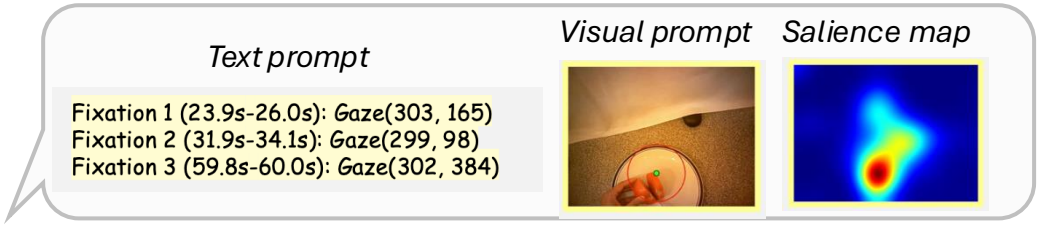
3. Results

- Across all gaze-conditioned streaming tasks, we highlight fundamental limits of current MLLMs. All MLLMs fall far below human performance. Models particularly struggle with temporal continuity, gaze grounding, and proactive prediction.

Method	Params	Frames	Past				Present				Proactive		Overall
			NFI	OTP	SR	GSM	OI (E)	OI (H)	OAR	FAP	GTA	OAA	
Human	-	-	0.700	0.889	0.903	0.707	0.960	0.920	0.800	0.840	0.765	0.780	0.827
Closed-Source MLLMs													
GPT-4o [17]	-	16	**0.601**	0.449	0.535	**0.580**	**0.729**	**0.730**	**0.596**	0.370	0.597	0.149	**0.535**
Claude Sonnet4 [18]	-	16	0.500	**0.554**	0.425	0.325	0.521	0.533	0.561	**0.439**	0.535	0.350	0.474
Claude Opus4 [18]	-	16	0.372	0.392	0.460	0.166	0.431	0.436	0.430	0.351	0.466	**0.490**	0.399
GazeQA-based Fine-tuned Models													
AssistGaze [7]	26M	32	0.294	0.131	0.310	0.294	0.278	0.250	0.109	0.254	N/A	N/A	0.223
Open-Source MLLMs													
Qwen2.5-VL [25]	7B	Adaptive	0.518	0.350	0.450	0.483	0.590	0.558	0.548	0.391	0.486	0.407	0.478
InternVL3.5 [26]	8B	Adaptive	0.490	0.311	**0.573**	0.548	0.627	0.628	0.466	0.372	0.373	0.051	0.444
VITA 1.5 [3]	7B	16	0.474	0.365	0.346	0.378	0.455	0.396	0.437	0.370	0.351	0.267	0.384
MiniCPM-V [32]	8B	32	0.430	0.374	0.354	0.296	0.334	0.379	0.438	0.345	0.480	0.216	0.365
Kangaroo [14]	7B	64	0.363	0.365	0.319	0.275	0.454	0.484	0.402	0.412	0.242	0.198	0.351
Open-Source Streaming MLLMs													
ViSpeak [4]	7B	1 fps	0.463	0.358	0.417	0.473	0.572	0.581	0.406	0.309	**0.635**	0.458	0.467
Dispider [21]	7B	1 fps	0.366	0.365	0.381	0.263	0.336	0.338	0.353	0.321	0.252	0.261	0.323
Flash-VStream [34]	7B	1 fps	0.249	0.202	0.336	0.220	0.289	0.147	0.044	0.280	0.443	0.217	0.243
VideoLLM-online [2]	8B	2 fps	0.000	0.000	0.000	0.000	0.006	0.006	0.002	0.000	0.458	0.333	0.080

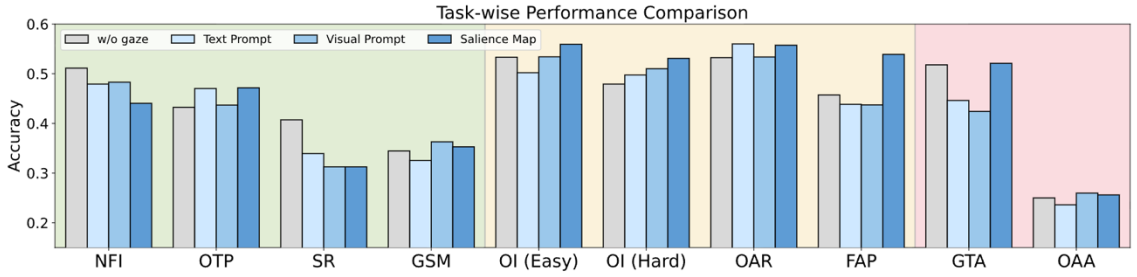


4. Additional Analysis



- **Ablation of gaze input prompting**: We analyze different strategies for incorporating gaze information into MLLMs

	Past	Present	Proactive	Avg.
Qwen2.5-VL [1] (w/o gaze)	**0.423**	0.500	0.384	0.446
+ test prompt	0.403	0.499	0.341	0.429
+ visual prompt	0.398	0.503	0.342	0.429
+ salience map	0.394	**0.546**	**0.386**	**0.454**



- **Fine-tuning results**: Due to the broad task coverage of StreamGaze, the resulting data distribution is inherently imbalanced, and performance improvements do not occur uniformly across all tasks.

Table 6. **Fine-tuning results on ViSpeak [4].** We construct four variants of training data for our fine-tuning experiments, drawing from both external resources [19, 20] and our own automatically generated data based on Ego4D-Gaze [5] and EgoExoLearn [6].

Fine-tuning datasets			Past				Present				Proactive		Overall
External [19, 20]	EgoExoLearn [6]	Ego4D-Gaze [5]	NFI	OTP	SR	GSM	OI (E)	OI (H)	OAR	FAP	GTA	OAA	
			0.526	0.486	0.341	0.418	0.635	0.477	0.413	0.489	0.504	0.502	0.479
✓			0.467	0.477	0.300	0.347	0.368	0.562	0.420	0.392	0.147	0.381	0.386
✓	✓		0.592	0.563	0.469	0.385	0.622	0.616	0.544	0.359	0.474	0.792	0.542
✓		✓	0.517	0.545	0.495	0.373	0.540	0.452	0.496	0.351	0.487	0.737	0.500
✓	✓	✓	0.583	0.540	0.495	0.388	0.601	0.641	0.575	0.331	0.409	0.651	0.521

NeurIPS 2026

Thank you!

Gaze-Guided Temporal Reasoning and Proactive Understanding in Streaming Videos

Daeun Lee, Subhojyoti Mukherjee, Branislav Kveton,
Ryan A. Rossi, Viet Dac Lai, Seunghyun Yoon, Trung Bui,
Franck Dernoncourt, Mohit Bansal

