NATIVEQUERY: BRINGING SPOKEN QUERIES INTO CONTINUOUS AUDIO-VISUAL STREAMS

Anonymous authors
Paper under double-blind review

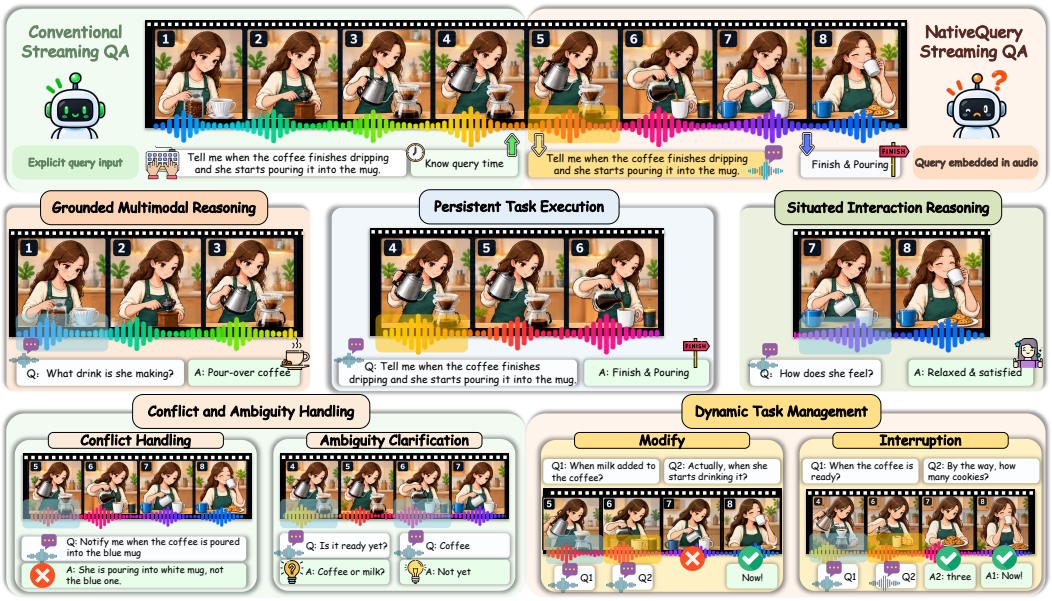


Figure 1: NativeQuery brings spoken queries into continuous audio-visual streams. Conventional streaming QA provides explicit queries at predefined timestamps, whereas NativeQuery embeds user requests directly into the ongoing audio stream and requires models to understand and respond to them in context.

ABSTRACT

Streaming omnimodal models can continuously perceive audio and visual inputs, but user queries are still typically provided separately as explicit inputs rather than arising within the stream itself. In real-world interactions, queries naturally occur as part of the continuous audio-visual stream, requiring models to recognize and interpret them while continuously processing the surrounding multimodal context. To address this challenge, we introduce **NativeQuery**, consisting of **NativeQuery-Train**, a large-scale training corpus, and **NativeQuery-Bench**, an evaluation benchmark for spoken-query interaction in continuous audio-visual streams. NativeQuery-Train contains 200K user-query turns, while NativeQuery-Bench contains 3,419 evaluation user-query turns across five task families and fifteen subtasks, covering grounded multimodal reasoning, conflict and ambiguity handling, persistent task execution, dynamic task management, and situated interaction reasoning. NativeQuery integrates queries into the original streams while preserving their native speakers and surrounding audio-visual context. We further introduce **NativeStream**, a streaming omnimodal model trained on NativeQuery-Train for such in-stream interaction. Experiments on NativeQuery-Bench reveal a substantial gap between explicit text queries and in-stream spoken queries, with Qwen 3.8 Omni Flash Realtime dropping from 64.5% to 42.8% in content accuracy. NativeStream substantially mitigates this problem, with NativeStream-7B achieving 36.2% joint accuracy on NativeQuery-Bench and 0.419 All-Global on OmniInteract, outperforming existing streaming omnimodal models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Real-time omnimodal models can continuously process audio and video, and increasingly support spoken interaction over live multimodal streams (Xu et al., 2025; Cui et al., 2026; Qwen Team, 2026b). However, existing streaming evaluations typically present user queries as explicit interaction turns, with both the query content and its timing directly specified to the model. This setup bypasses a fundamental challenge of natural interaction: the user query may itself occur within the ongoing multimodal stream. For example, a request such as *tell me when the water in the pan starts boiling* may be embedded in the same audio stream as scene speech and environmental sounds. The model must recognize the utterance as a user request, ground it in the concurrent visual and acoustic context, maintain the request as the scene evolves, and update or cancel it when the user's intent changes. The query therefore serves not only as an instruction about the stream, but also as part of the stream that the model must perceive and interpret, as illustrated in Fig. 1.

Several recent studies explore spoken queries in omnimodal interaction. OmniAssistBench evaluates audio-visual assistant interaction on pre-constructed clips, but adopts offline inference rather than native streaming inference (Sun et al., 2026). ROMA moves toward online interaction by incorporating spoken queries into streaming inputs (Tian et al., 2026). Its benchmark, however, is largely built from repurposed vision-centric datasets, and the inserted queries are synthesized with voices that differ from those in the original videos, making the queries acoustically distinct from the surrounding stream. OmniInteract is more closely aligned with our setting, preserving continuous audio-visual input and evaluating models through native online inference (Lu et al., 2026b). Its coverage, however, is concentrated on Chinese daily-life interactions with relatively limited cross-modal reasoning, while its smaller English subset primarily consists of instructional videos. Moreover, user queries are typically placed near the beginning of the video rather than occurring at diverse points throughout the stream. As a result, existing benchmarks provide limited coverage of spoken queries that emerge naturally at different times and require reasoning over evolving audio-visual context.

To address this gap, we introduce **NativeQuery**, a unified train-and-evaluate resource comprising **NativeQuery-Train** and **NativeQuery-Bench**. NativeQuery-Train contains 200K spoken user-query turns, while NativeQuery-Bench contains 3,419 evaluation user-query turns. For each source video, we extract temporally localized multimodal evidence, including visual entities and actions, visible text, speech, and non-speech sounds. We then link these observations through semantic and temporal relations. These relations guide task-specific query construction, determining both what the user asks and when the query occurs in the stream. Each query is verified against its supporting source evidence and then synthesized with a voice matched to the source and integrated into the original audio track. This preserves the surrounding audio-visual context. The resulting annotations capture query timing and temporally localized supporting evidence.

Experiments reveal a substantial gap between explicit and in-stream queries. Under matched query content, Qwen3.8-Omni-Flash Realtime (Qwen Team, 2026b) achieves 64.5% content accuracy with text queries, but only 42.8% when the same queries must be recognized from the audio stream. We further train **NativeStream** on NativeQuery-Train to examine whether training on in-stream spoken queries can reduce this gap. NativeStream-7B narrows the text-to-stream gap to 3.6 points, achieving 66.8% and 63.2% content accuracy under the two settings. On NativeQuery-Bench, it reaches 36.2% joint content-and-timing accuracy, compared with 26.7% for MiniCPM-o-4.5-9B (Cui et al., 2026). The improvement also transfers beyond NativeQuery, increasing the OmniInteract All-Global score from 0.368 to 0.419.

The main contributions are summarized as follows:

- We introduce **NativeQuery**, a large-scale dataset suite for training and evaluating spoken-query interaction in continuous audio-visual streams. It covers five task families and fifteen subtasks spanning grounded multimodal reasoning, conflict and ambiguity handling, persistent task execution, dynamic task management, and situated interaction reasoning.
- We develop an evidence-driven construction pipeline that extracts and connects temporally localized visual, speech, sound, and text evidence to generate task-specific queries.
- We introduce **NativeStream**, a streaming omnimodal model trained on **NativeQuery-Train**, as a baseline for in-stream spoken-query interaction. Experiments on NativeQuery-Bench and OmniInteract validate the effectiveness of training on NativeQuery-Train.

2 Related Work

2.1 Omnimodal Benchmarks

Early audio-visual benchmarks evaluate reasoning over complete clips with externally provided questions (Alamri et al., 2019; Li et al., 2022; 2026). Recent benchmarks move toward online interaction: StreamingBench, OmniMMI, LiViBench, and SocialOmni introduce timestamped questions, interaction history, livestream content, and social turn-taking (Lin et al., 2026; Wang et al., 2025b; 2026; Xie et al., 2026b), while PhoStream, OmniPro, OmniDuplexEval, and StreamArena explicitly evaluate evidence availability or response timing (Lu et al., 2026a; Zhao et al., 2026; He et al., 2026; Zhang et al., 2026). These benchmarks establish the importance of causal evidence and timely responses, but generally treat the query as a separately delivered input rather than an event that must itself be recognized within the continuous stream.

The closest resources differ primarily in how streaming interactions are constructed and supervised. OmniAssistBench uses edited and segmented videos to simulate assistant-style multi-turn interaction offline (Sun et al., 2026). ROMA reformulates existing datasets into streaming training examples covering proactive alerts, narration, and reactive QA (Tian et al., 2026). OmniInteract evaluates continuous interaction through temporally grounded response slots, each annotated with a trigger, response window, and target answer (Lu et al., 2026b). In contrast, NativeQuery places spoken requests directly within continuous audio-visual streams and associates the corresponding responses with timestamped multimodal evidence and interaction events. NativeQuery further provides matched training and evaluation data under a unified task design, enabling systematic study of in-stream spoken-query interaction.

2.2 Streaming Omnimodal Understanding

Streaming omnimodal models increasingly support joint audio-visual perception, speech generation, interruption handling, and multi-turn dialogue (Fu et al., 2024; Xie & Wu, 2024; Tong et al., 2025; Xu et al., 2025; Cui et al., 2026; Tian et al., 2026). To operate over long streams, existing systems maintain compact or hierarchical memories and update visual context online (Zhang et al., 2024a; Liu et al., 2024; Yao et al., 2025; Zhang et al., 2024b; Xie et al., 2026a). Other approaches further learn when sufficient evidence has become available and whether a response should be generated at the current time (Chen et al., 2024; Wang et al., 2025a; Azad et al., 2026; Zhang et al., 2026). These works improve the ability of models to continuously perceive, remember, and respond to streaming inputs, but they generally assume that user queries are provided as explicit inputs. As a result, interaction settings in which the query itself is part of the continuous audio-visual stream and must be handled within that stream remain underexplored.

Existing training methods mainly supervise behaviors such as online response generation, silence, proactive alerts, or continuous narration (Chen et al., 2024; Xia et al., 2026; Tian et al., 2026). NativeQuery-Train instead provides continuous audio-visual training samples containing in-stream spoken requests, together with corresponding response supervision across different interaction scenarios. We train NativeStream on NativeQuery-Train to study whether such training data can improve a model's ability to handle spoken requests embedded within audio-visual streams.

3 NativeQuery

3.1 Task Definition

NativeQuery studies spoken user queries embedded within continuous audio-visual streams. Each episode follows a causal online protocol: at time t, the model has access only to the stream prefix and interaction history observed up to t. Because supporting evidence may precede, overlap with, or follow the query, the model must determine both what response is justified and when that response becomes valid. We organize this setting into five task families and 15 subtasks, illustrated in Fig. 2, and use the abbreviations defined below throughout the paper.

A. Grounded Multimodal Reasoning (GrdRe). Grounded Multimodal Reasoning requires each response to be traceable to relevant evidence in the observed stream, rather than inferred from the query wording or model priors alone. Its five subtasks vary the evidence to be located and the reasoning performed over it.



Figure 2: **Five task families in NativeQuery.** All tasks embed spoken queries into continuous audio-visual streams, while different families target multimodal grounding, ambiguity handling, persistent execution, task-state management, and situated interaction reasoning.

- *Text Grounding* (*TxtGrd*). Identify scene text relevant to the spoken request and interpret it together with its visual carrier and surrounding scene.
- *Referential Grounding* (*RefGrd*). Resolve a spoken referring expression to the intended person, object, or speaker using spatial, visual, temporal, and speaker-identity cues.
- *Event Grounding* (*EvtGrd*). Identify and temporally localize an action, state change, or sound-producing event referenced by the query or indicated by an acoustic cue.
- *Relational Reasoning* (*RelRea*). Derive spatial, quantitative, comparative, causal, or cross-modal relations among grounded entities and events.
- *Temporal Reasoning* (*TemRea*). Reason about event order, duration, repetition, state transitions, and other temporal dependencies using evidence distributed across the stream.

B. Conflict and Ambiguity Handling (ConAmb). Conflict and Ambiguity Handling covers cases in which the observed evidence does not justify literal compliance with the request or commitment to a single interpretation. It separates resolvable false premises from genuine ambiguity that requires another user turn.

- *Conflict Resolution* (*ConRes*). Identify a premise contradicted by the observed stream and correct it with an evidence-supported account.
- *Ambiguity Clarification* (*AmbClar*). When the context supports multiple plausible interpretations, ask a targeted question that distinguishes among them instead of selecting one without sufficient evidence.

C. Persistent Task Execution (PerExe). Persistent Task Execution covers standing requests whose completion depends on future observations. The request remains unchanged while the model updates task-relevant state as new observations arrive and responds only when the specified trigger or reporting condition is satisfied.

- *Event Monitoring* (*EvtMon*). Maintain a pending request until one or more specified future events occur, responding whenever a trigger condition is met.
- *Progress Tracking* (*ProgTrk*). Accumulate task state toward a finite goal and report the requested intermediate or final status.
- *Procedure Supervision* (*ProcSup*). Track completed, current, and pending steps under an expected order, guide the next action, and flag omissions or out-of-order execution.

D. Dynamic Task Management (DynMan). Dynamic Task Management focuses on changes introduced by subsequent user turns. Unlike **PerExe**, where the request remains fixed while the scene evolves, these tasks require the model to revise the active obligation or temporarily prioritize a new request without discarding a still-valid goal.

- *Request Revision* (*ReqRev*). Incorporate a correction, replacement, or other modification into the active request and suppress response events that remain valid only under the superseded version.

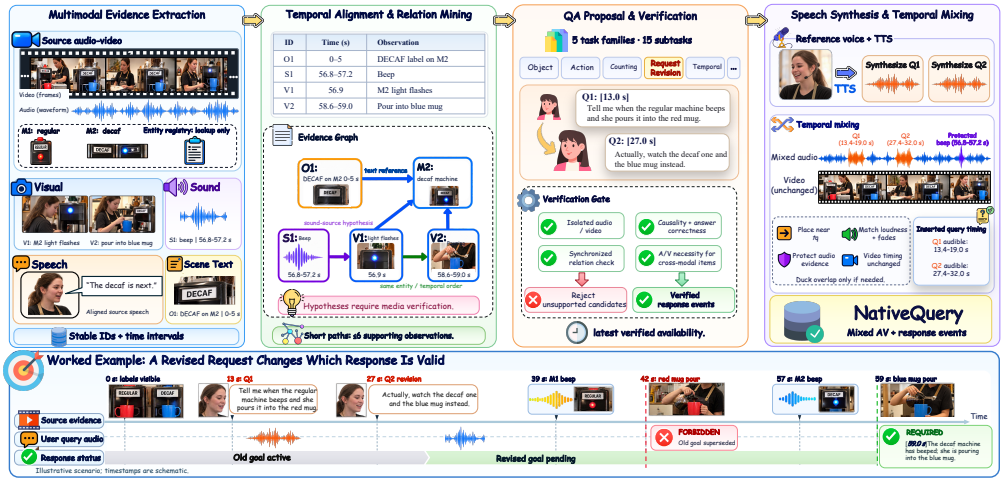


Figure 3: Data construction pipeline for NativeQuery.

- *Interruption and Resumption* (*IntrRes*). Address an intervening query while preserving the original request, then resume the original task without requiring the user to repeat it.

E. Situated Interaction Reasoning (SitInt). Situated Interaction Reasoning addresses meaning that cannot be recovered from an utterance transcript alone. Here the model must combine lexical content with prosody, visible behavior, participant reactions, and interaction history to interpret the ongoing social exchange.

- *Affective Stance* (*AffSta*). Infer a participant's affective or attitudinal orientation toward a situation, another participant, or an ongoing event.
- *Pragmatic Intent* (*PragInt*). Recover the intended communicative function of an utterance or behavior from the surrounding multimodal context.
- *Interaction Dynamics* (*IntDyn*). Track how participant roles, turn-taking, engagement, and reactions change over the course of an interaction.

Together, the five families cover evidence-grounded reasoning and conflict or ambiguity handling (**GrdRe**, **ConAmb**), persistent and evolving task states (**PerExe**, **DynMan**), and socially situated interpretation (**SitInt**); all require both correct content and temporally valid responses.

3.2 Dataset Construction

As shown in Fig. 3, we construct each episode in four stages that preserve temporal provenance while distinguishing the authored query time t_q, the evidence-sufficiency time t_{suf}, and the realized query-audio boundaries.

Stage 1. Multimodal evidence extraction. We first separate each source video into visual and audio streams. From the visual stream, we build an entity registry of distinguishable people, animals, and objects, each with a stable identifier and visible intervals; the registry supports lookup and reference resolution but does not constitute evidence. From the audio stream, we extract non-speech sounds and force-aligned speech, assign stable speaker identities, and link voices to visible entities only with synchronized audio-visual evidence. We also annotate visible actions, interactions, state changes, and legible scene text with its carrier, information type, and readable interval, excluding subtitles and closed captions because they duplicate speech. Each timeline observation stores a modality, description, stable reference, and temporal span.

Stage 2. Temporal alignment and relation mining. We merge all observations into a normalized timeline, converting coarse annotations to bounded intervals while preserving source references. We then build a typed evidence graph over speech, sound, visual events, and scene text. Candidate edges encode sound source, speaker binding, speech or text reference, same entity, state transition, claim observation, and temporal order; each carries a specific claim, supporting references, and a verification question. These edges remain authoring hypotheses rather than facts inferred from temporal overlap. Bounded graph traversal mines short cross-modal paths with at most six supporting ob-

servations while preserving relation identifiers and evidence references. Entity descriptions remain separate lookup hints and cannot silently become evidence for an action or identity.

Stage 3. QA proposal and verification. For each task family and subtask, a task-conditioned generator selects a suitable evidence path and proposes a spoken query, a candidate response, a query time t_q, and a compact evidence set. It records the facts known at t_q and the information still unresolved; each candidate must express one coherent goal and cite one to six non-redundant observations. The wording must obey streaming causality: it cannot state or presuppose a fact that is unavailable when the query begins, although the response may depend on evidence that arrives later. For a cross-modal candidate, removing all audio evidence or all visual evidence must leave the response unresolved or non-unique.

Each proposal then undergoes media-grounded verification against the original media. We check cited claims in modality-isolated clips, estimate their earliest verified availability, verify graph relations in synchronized audio-visual clips, and conduct an independent semantic review of response correctness, uniqueness, atomicity, future leakage, modality necessity, and evidence redundancy. Unsupported or ambiguous candidates are discarded, as are cross-modal candidates for which either modality is unnecessary; the pipeline does not repair them by rewriting claims or pruning evidence. For a retained item with evidence set E and selected relations P, we record

$$t_{\text{suf}} = \max(\{a_e : e \in E\} \cup \{a_r : r \in P\}) . \tag{1}$$

Here a_e is the verified availability time of observation e, while a_r is the upper end of the verified availability interval for relation r. Thus, t_{suf} marks the completion of the documented evidence chain and may occur before or after t_q. We convert retained candidates into ordered response events linked to supporting evidence, target time, and validity. A standing request can produce multiple responses; if a later turn replaces it, responses tied only to the superseded request must be suppressed while the scene evidence remains valid, separating changes in user intent from changes in the media.

Stage 4. Speech synthesis and temporal mixing. We synthesize each user query using reference clips sampled from diarized source speech when available and from ASR-aligned speech segments otherwise. The generated clip is checked for runaway generation and trimmed to its active speech region. The mixer places each turn near t_q using local audio quietness, ASR speech intervals, speaker boundaries, visual event boundaries, protected audio evidence spans, and the positions of earlier turns. It first seeks a clean window that avoids transcribed speech, protected audio evidence, and other queries; if none exists, it enlarges the search region and selects the quietest low-overlap window while prioritizing protected evidence, relaxing constraints only when no legal window remains. The TTS level is matched to the local scene loudness, and short fades reduce abrupt transitions. When overlap with source speech is unavoidable, the original mixed track is smoothly attenuated, which may also affect ambient sound because the source has a single mixed track. The video stream is copied without temporal editing. The final annotations store the actual audible query boundaries separately from the authored time t_q. Multiple turns from one record share a mixed video, whereas records that reuse a source video are rendered independently to prevent cross-record contamination.

3.3 NativeQuery-Bench Statistics

Figure 4 summarizes the NativeQuery-Bench, which contains 3,419 user query turns embedded in 2,050 distinct source videos and covers five task families and 15 subtasks. The source videos last 56–180 seconds (median: 103 seconds), with a cumulative duration of 59.9 hours. Across these user query turns, we provide 4,331 response annotations, including 320 superseded response events following request revisions. These annotations cite 9,221 evidence items: 3,524 visual observations, 3,388 speech segments, 1,612 scene text items, and 697 non-speech sound events.

Evidence and response timing vary substantially across task families. For **GrdRe** and **ConAmb**, cited evidence midpoints precede query end by medians of 15.4 and 9.1 seconds, respectively. For **PerExe** and **DynMan**, they follow query end by 15.6 and 8.0 seconds. The median interval from query end to the annotated target time is 0 seconds for **GrdRe** and **ConAmb**, 20.0 seconds for **PerExe**, 8.5 seconds for **DynMan**, and 4.2 seconds for **SitInt**. Within **PerExe**, 600 user query turns yield 1,512 response annotations because a standing request can trigger multiple updates. These patterns contrast retrospective reasoning with interactions that depend on later evidence or updated user intent.

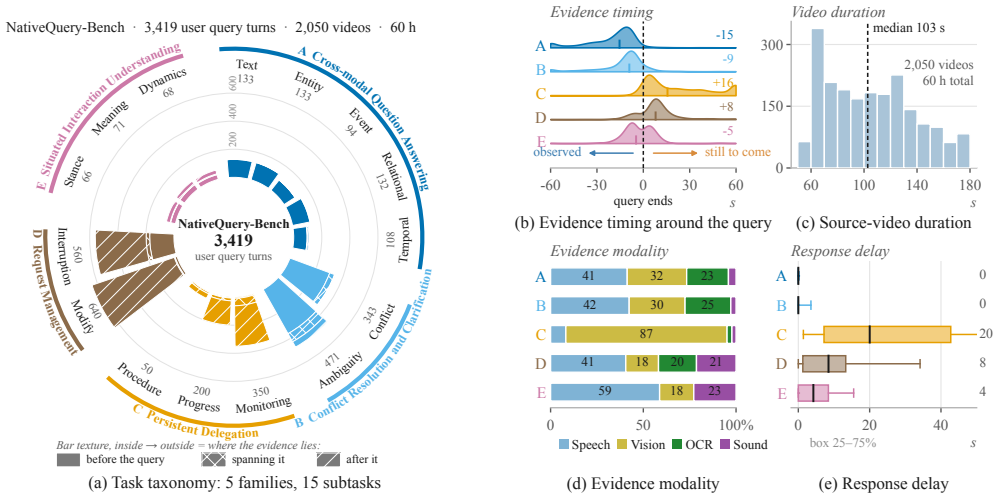


Figure 4: NativeQuery-Bench statistics. The 3,419 user query turns yield 4,331 response annotations. (a) Annotation counts by task family and subtask. The radial segments indicate whether cited evidence ends before, straddles, or begins after query end. (b) Evidence-span midpoints relative to query end. (c) Distribution of source-video durations. (d) Evidence-modality proportions by task family. (e) Interval from query end to the annotated target time.

4 NativeStream

NativeStream is a streaming omnimodal model initialized from Qwen2.5-Omni and fine-tuned on NativeQuery-Train to handle spoken requests embedded in continuous audio-visual streams. We keep spoken requests in the audio and mark each annotated answer time with `</Response>` followed by the target answer (Xia et al., 2026). Other Qwen2.5-Omni model settings follow the default configuration (Xu et al., 2025).

5 Experiments

5.1 Experimental Setup

Evaluated Models. On NativeQuery-Bench, we compare NativeStream-3B and NativeStream-7B with Qwen 3.0-Omni Flash Realtime (Qwen Team, 2025), Qwen 3.5-Omni Plus Realtime (Qwen Team, 2026a), Qwen 3.8-Omni Flash Realtime (Qwen Team, 2026b), Gemini 2.5 Flash Live, Gemini 3.8 Live, ROMA-7B (Tian et al., 2026), and MiniCPM-o-4.5-9B (Cui et al., 2026). The query-modality study additionally includes VITA-1.5-7B (Fu et al., 2025), Qwen2.5-Omni-3B and Qwen2.5-Omni-7B (Xu et al., 2025), Qwen3-Omni-30B-A3B (Qwen Team, 2025), Gemini 3.8 Flash, and Gemini 3.1 Pro. For transfer on OmniInteract (Lu et al., 2026b), we evaluate AURA (Lu et al., 2026c), Gemini 2.5 Flash Live, MiniCPM-o-4.5-9B, Qwen3.5-Omni Flash Realtime (Qwen Team, 2026a), and both NativeStream models.

Implementation Details. NativeQuery draws source videos from OmniVideo-100K (Cai et al., 2026), COIN (Tang et al., 2019), and YouCook2 (Zhou et al., 2018). During data construction, Gemini 3.7 Flash extracts multimodal evidence, while GPT-5.6-sol proposes QA pairs and performs text-based verification. We transcribe and align speech with Qwen3-ASR-1.7B-hf and Qwen3-ForcedAligner-0.6B-hf (Shi et al., 2026), and synthesize spoken queries with Qwen3-TTS-12Hz-1.7B-Base (Hu et al., 2026). We fine-tune NativeStream-3B and NativeStream-7B from Qwen2.5-Omni (Xu et al., 2025) on NativeQuery-Train for one epoch using autoregressive response-time supervision (Xia et al., 2026). On NativeQuery-Bench, models receive only the observed audio-visual prefix and interaction history; GPT-5.6-sol judges response content, and responses within three seconds of a target are considered timely. For Request Revision, the superseded response is scored as a forbidden target and the revised response as required; for Interruption and Resumption, the intervening response and the resumed original-task response are both required.

Table 1: Performance comparison on StreamingOmni. For each task family (GrdRe, ConAmb, PerExe, DynMan, and SitInt; Section 3.1) and the overall column we report response-timing accuracy (T), content accuracy (C), and joint accuracy (J). **Overall** is the macro average across the five task families.

Model	GrdRe			ConAmb			PerExe			DynMan			SitInt			Overall		
	T↑	C↑	J↑	T↑	C↑	J↑	T↑	C↑	J↑	T↑	C↑	J↑	T↑	C↑	J↑	T↑	C↑	J↑
Proprietary Streaming Omni Models																		
Qwen 3.0-Omni Flash R	39.4	42.7	16.8	37.1	27.2	10.1	61.8	59.4	36.7	40.7	20.6	8.4	52.9	57.8	30.6	46.4	41.5	20.5
Qwen 3.5-Omni Plus R	35.6	61.7	22.0	33.8	36.9	12.5	63.1	68.4	**43.2**	36.7	30.8	11.3	54.0	63.1	**34.1**	44.6	52.2	24.6
Qwen 3.8-Omni Flash R	36.8	48.9	18.0	35.0	29.5	10.3	**66.5**	56.8	37.8	38.5	23.3	9.0	56.1	55.6	31.2	46.6	42.8	21.3
Gemini 2.5 Flash Live	42.8	49.4	21.1	38.6	31.8	12.3	38.9	42.7	16.6	36.2	14.8	5.4	47.6	42.1	20.0	40.8	36.2	15.1
Gemini 3.8 Live	40.5	52.7	21.3	35.3	35.9	12.7	35.4	46.1	16.3	34.0	16.7	5.7	50.2	44.7	22.4	39.1	39.2	15.7
Open-source Streaming Omni Models																		
ROMA-7B	44.7	10.8	4.8	42.1	7.6	3.2	64.9	16.3	10.6	40.3	6.5	2.6	**57.6**	8.5	4.9	49.9	9.9	5.2
MiniCPM-o-4.5-9B	67.5	65.9	44.5	56.5	34.8	19.7	44.8	84.5	37.8	28.2	28.6	8.1	39.0	60.0	23.4	47.2	54.8	26.7
NativeStream-3B	63.8	61.9	39.5	65.2	45.8	29.9	48.7	77.1	37.5	45.1	43.4	19.6	42.9	61.5	26.4	53.1	57.9	30.6
NativeStream-7B	**68.0**	**66.5**	**45.2**	**70.1**	**52.3**	**36.7**	52.6	**82.0**	43.1	**52.0**	**48.6**	**25.3**	46.3	**66.4**	30.7	**57.8**	**63.2**	**36.2**

5.2 Main Results

Performance on NativeQuery.

Table 1 reveals a substantial gap between answering correctly and responding at the right time. The nine billion parameter MiniCPM model achieves 54.8% overall content accuracy, but its joint accuracy is only 26.7%. This gap shows that existing streaming omni models still struggle to coordinate response content with response timing. Our three billion parameter model improves these scores from 47.2%, 54.8%, and 26.7% to 53.1%, 57.9%, and 30.6%, respectively. Our seven billion parameter model further raises these scores to 57.8%, 63.2%, and 36.2%. These results indicate that supervision from NativeQuery enables a smaller model to outperform a larger baseline across all three evaluation dimensions.

Joint accuracy offers a stringent measure of streaming interaction because a response is correct only when both its content and timing are valid. Our seven billion parameter model achieves the highest joint accuracy in all five task families. The largest gains occur in ConAmb and DynMan. Joint accuracy increases from 19.7% to 36.7% in ConAmb and from 8.1% to 25.3% in DynMan relative to the MiniCPM model. These findings show that NativeQuery captures capabilities that content only evaluation misses. Its supervision aligns responses with available evidence and with the evolving interaction state. This alignment is essential for reliable streaming interaction.

Generalization to OmniInteract. To assess transfer beyond the NativeQuery task taxonomy, we evaluate the models on OmniInteract with the IA-QTF1 metric. OmniInteract covers real time, proactive, nested, and 1QnA interactions. NativeStream-7B obtains the best 1Q1A Global score of 0.527 and the best All Global score of 0.419. This exceeds AURA by 0.060 on 1Q1A Global and MiniCPM-o-4.5-9B by 0.051 on All Global. The model is particularly strong in proactive and nested interactions. It reaches 0.620 and 0.610 in these settings, exceeding MiniCPM-o-4.5-9B by 0.013 and 0.011. Because OmniInteract uses a different task organization, these gains suggest that the temporal state tracking learned from NativeQuery transfers beyond its original task definitions. The real time setting remains more difficult. NativeStream-7B obtains 0.450, below Gemini 2.5 Flash Live at 0.553 and Qwen3.5-Omni Flash Realtime at 0.524. This gap suggests that low latency response timing remains a harder aspect of transfer. The 1QnA setting is difficult for all models. The best score is only 0.059, achieved by NativeStream-7B. Scaling from 3B to 7B improves every reported setting, with All Global increasing from 0.356 to 0.419. The consistent gains across settings indicate that the transfer benefit is not limited to a single interaction type.

5.3 Ablation Analysis

Effect of Query Modality. Table 3 compares matched queries in audio and text form. Text queries usually produce higher content accuracy. The size of this advantage varies substantially across models. In Oracle Mode, text raises the overall content accuracy of Qwen 2.5 Omni 3B from 26.6

Table 2: IA-QTF1 across interaction settings. The 1Q1A columns use mutually exclusive real-time, proactive, and nested response slots; global scores are recomputed from aggregated TP/FP/FN.

Model	1Q1A				1QnA	All Global
	Real-time	Proactive	Nested	1Q1A Global		
AURA	0.376	0.549	0.596	0.467	0.052	0.363
Gemini 2.5 Flash Live	**0.553**	0.121	0.398	0.428	0.028	0.344
MiniCPM-o-4.5-9B	0.337	0.607	0.599	0.456	0.015	0.368
Qwen3.5-Omni Flash Realtime	0.524	0.108	0.379	0.401	0.023	0.323
NativeStream-3B	0.361	0.558	0.552	0.450	0.047	0.356
NativeStream-7B	0.450	**0.620**	**0.610**	**0.527**	**0.059**	**0.419**

Table 3: **Content accuracy (C_Acc, %) on NativeQuery-Bench.** We compare audio and text queries under Oracle and Streaming modes across five task families: grounded multimodal reasoning (GrdRe), conflict and ambiguity handling (ConAmb), persistent task execution (PerExe), dynamic task management (DynMan), and situated interaction reasoning (SitInt). Overall denotes the macro average across the five task families.

Method	Query Type	GrdRe	ConAmb	PerExe	DynMan	SitInt	Overall
Oracle Mode							
VITA-1.5-7B	Audio	37.0	23.7	13.2	4.6	19.5	19.6
	Text	32.7	26.6	34.7	10.0	33.8	27.6
Qwen 2.5-Omni-3B	Audio	33.2	19.4	40.0	12.5	27.9	26.6
	Text	66.1	36.3	80.2	44.0	68.7	59.1
Qwen 2.5-Omni-7B	Audio	50.4	25.1	46.7	20.4	40.8	36.7
	Text	67.9	41.4	84.0	46.5	70.2	62.0
Qwen 3-Omni-30B-A3B	Audio	19.1	14.9	58.2	15.2	16.1	24.7
	Text	61.3	42.8	87.2	29.2	67.3	57.6
Gemini 3.8 Flash	Audio	82.8	65.0	87.1	38.5	77.6	70.2
	Text	86.9	64.0	90.2	46.5	86.6	74.8
Gemini 3.1 Pro	Audio	85.1	68.6	90.8	42.1	82.7	73.9
	Text	89.2	72.4	94.1	50.8	89.3	79.2
Streaming Mode							
MiniCPM-o-4.5-9B	Audio	65.9	34.8	84.5	28.6	60.0	54.8
	Text	68.7	39.6	85.9	31.3	68.4	58.8
Qwen 3.8-Omni Flash Realtime	Audio	48.9	29.5	56.8	23.3	55.6	42.8
	Text	78.2	51.9	73.2	30.3	89.2	64.5
Gemini 3.8 Live	Audio	52.7	35.9	46.1	16.7	44.7	39.2
	Text	64.9	43.7	55.8	22.4	61.5	49.7
NativeStream-7B	Audio	66.5	52.3	82.0	48.6	66.4	63.2
	Text	69.8	57.6	84.1	52.4	70.3	66.8

percent to 59.1 percent. For Qwen 3 Omni 30B A3B, it raises the score from 24.7 percent to 57.6 percent. The corresponding changes are smaller for Gemini 3.8 Flash, from 70.2 percent to 74.8 percent, and Gemini 3.1 Pro, from 73.9 percent to 79.2 percent. These results suggest that stronger multimodal models extract more useful information from audio queries, although text remains the more favorable query format. Our 7B model shows a smaller dependence on query modality in Streaming Mode. It reaches 63.2 percent with audio queries and 66.8 percent with text queries, the smallest separation among the streaming models. Across the five task families, the two settings remain within 5.3 of each other. This consistency indicates that the NativeQuery training design helps the model preserve query semantics while interpreting continuous multimodal input in real time. It also supports natural spoken interaction without requiring a separately prepared text query. The occasional reversals, such as the audio advantage of VITA 1.5 7B on GrdRe and Gemini 3.8 Flash on ConAmb, show that the effect of query modality depends on task requirements. Overall, our 7B model demonstrates a more balanced response to audio and text queries.

6 Conclusion

We introduced NativeQuery, comprising NativeQuery-Train and NativeQuery-Bench, and NativeStream for spoken-query interaction in continuous audio-visual streams. NativeQuery-Bench evaluates whether models recognize and ground in-stream requests and respond with correct content at the appropriate time. NativeStream-7B achieves 36.2% joint content-and-timing accuracy on NativeQuery-Bench and 0.419 All Global on OmniInteract, demonstrating the value of training on NativeQuery-Train. Future work will replace synthesized queries with natural speech and improve low-latency response triggering in longer real-world interactions.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, including improving clarity, grammar, and presentation. All AI-assisted text was reviewed and revised by the authors to ensure technical accuracy and consistency. The authors take full responsibility for the final content of this paper.

REFERENCES

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio Visual Scene-Aware Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7550–7559, 2019.

Shehreen Azad, Vibhav Vineet, and Yogesh S. Rawat. StreamReady: Learning What to Answer and When in Long Streaming Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. OmniVideo-100K: A Dataset for Audio-Visual Reasoning through Structured Scripts and Evidence Chains. *arXiv preprint arXiv:2606.14702*, 2026.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. 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.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. MiniCPM-o 4.5: Towards Real-Time Full-Duplex Omni-Modal Interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, Haoyu Cao, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu et al. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Chaoqun He, Mingyang Xiang, Yingjing Xu, Bokai Xu, Junbo Cui, Jie Zhou, Yuan Yao, and Lijie Wen. Omni-DuplexEval: Evaluating Real-time Duplex Omni-modal Interaction. *arXiv preprint arXiv:2605.17360*, 2026.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-TTS Technical Report. *arXiv preprint arXiv:2601.15621*, 2026.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding. In *ICASSP 2026–2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151, 2026.

Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M. Alvare. StreamChat: Chatting with Streaming Video. *arXiv preprint arXiv:2412.08646*, 2024.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Li, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. PhoStream: Benchmarking Real-World Streaming for Omnimodal Assistants in Mobile Scenarios. *arXiv preprint arXiv:2601.22575*, 2026a.

Xudong Lu, Xueying Li, Annan Wang, Yang Bo, Jinpeng Chen, Zengliang Li, Nianzu Yang, Rui Liu, Xue Yang, Jingwen Hou, and Hongsheng Li. OmniInteract: Benchmarking Real-World Streaming Interaction for Real-Time Omnimodal Assistants. *arXiv preprint arXiv:2605.26485*, 2026b.

Xudong Lu et al. AURA: Always-On Understanding and Real-Time Assistance via Video Streams. *arXiv preprint arXiv:2604.04184*, 2026c.

Qwen Team. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765*, 2025.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026a.

Qwen Team. Qwen3.8-Omni: Towards Native Omni-Modal Agents. *arXiv preprint arXiv:2609.25611*, 2026b.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR Technical Report. *arXiv preprint arXiv:2601.21337*, 2026.

Xianyun Sun, Chaoyou Fu, Zhengye Zhang, Feiyang Duan, Qingyuan Cao, Yonghui Niu, Sihang Yuan, Ge Zhang, and Caifeng Shan. OmniAssistBench: Assistant-style Interaction Benchmark for Omni-LLMs. *arXiv preprint arXiv:2608.21360*, 2026.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. COIN: A Large-Scale Dataset for Comprehensive Instructional Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Xueyun Tian, Wei Li, Bingbing Xu, Heng Dong, Yuanzhuo Wang, and Huawei Shen. ROMA: Real-time Omni-Multimodal Assistant with Interactive Streaming Understanding. *arXiv preprint arXiv:2601.10323*, 2026.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. InteractiveOmni: A Unified Omni-modal Model for Audio-Visual Multi-turn Dialogue. *arXiv preprint arXiv:2510.13747*, 2025.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. StreamBridge: Turning Your Offline Video Large Language Model into a Proactive Streaming Assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2025a.

Xiaodong Wang, Langling Huang, Zhirong Wu, Xu Zhao, Teng Xu, Xuhong Xia, and Peixi Peng. LiViBench: An Omnimodal Benchmark for Interactive Livestream Video Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(31):26517–26525, 2026.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. OmniMMI: A Comprehensive Multi-modal Interaction Benchmark in Streaming Video Contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18925–18935, 2025b.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming Video Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Ming Xie, Zizheng Huang, Xudong Tan, Chao Wang, Xiangyu Zeng, Wenxiao Wu, Tao Chen, Limin Wang, and Yanwei Fu. StreamOV: Streaming Omni-Video Understanding via Evidence-Guided Memory and Response Triggering. *arXiv preprint arXiv:2605.25621*, 2026a.

Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, Ruize Fang, Jiebo Luo, Rongrong Ji, and Xiawu Zheng. SocialOmni: Benchmarking Audio-Visual Social Interactivity in Omni Models. *arXiv preprint arXiv:2603.16859*, 2026b.

Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities. *arXiv preprint arXiv:2410.11190*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. 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.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-VStream: Memory-Based Real-Time Understanding for Long Video Streams. *arXiv preprint arXiv:2406.08085*, 2024a.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions. *arXiv preprint arXiv:2412.09596*, 2024b.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. StreamArena: Toward Continuous, Interactive, and Long-Horizon Agentic Streaming Video Understanding. *arXiv preprint arXiv:2608.05703*, 2026.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. OmniPro: A Comprehensive Benchmark for Omni-Proactive Streaming Video Understanding. *arXiv preprint arXiv:2605.18577*, 2026.

Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards Automatic Learning of Procedures from Web Instructional Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

This appendix is organized as follows:

- NativeQuery-Train Statistics (Section A).
- Response-Triggering Analysis (Section B).
- Data Synthesis Prompts (Section C).
- Limitations and Future Work (Section D).

A NativeQuery-Train Statistics

Table 4 summarizes NativeQuery-Train at distinct units of analysis. Its 159,998 dialogue records contain 224,161 spoken user-query turns and 281,551 response annotations, including 263,315 required and 18,236 forbidden annotations, associated with 11,162 distinct source videos. A record can contain multiple user turns or multiple response annotations. Video counts are deduplicated within each row and across the full split in the total row, so task-family video counts are not additive. Evidence counts refer to citations attached to response annotations; an observation cited by multiple events contributes once per event.

Table 4: NativeQuery-Train scale by task family and subtask. A record is one dialogue example; turns are spoken user requests, and annotations are required or forbidden response events. Evidence counts cited observations, including reuse across events. Videos are distinct source videos within each row; only the total row deduplicates across families. Req. and Forb. denote required and forbidden annotations.

Task family / subtask	Records	Videos	Turns	Annotations	Evidence	Req.	Forb.
GrdRe	39,999	8,507	39,999	39,999	83,381	39,999	0
Text Grounding (TxtGrd)	8,000	4,024	8,000	8,000	16,342	8,000	0
Referential Grounding (RefGrd)	8,000	4,002	8,000	8,000	18,596	8,000	0
Event Grounding (EvtGrd)	7,999	3,400	7,999	7,999	16,165	7,999	0
Relational Reasoning (RelRea)	8,000	4,335	8,000	8,000	16,144	8,000	0
Temporal Reasoning (TemRea)	8,000	3,284	8,000	8,000	16,134	8,000	0
ConAmb	30,980	4,087	58,671	58,671	91,014	58,671	0
Conflict Resolution (ConRes)	2,327	2,070	2,327	2,327	6,014	2,327	0
Ambiguity Clarification (AmbClar)	28,653	3,029	56,344	56,344	85,000	56,344	0
PerExe	35,484	5,263	35,484	92,874	92,874	92,874	0
Event Monitoring (EvtMon)	13,363	4,172	13,363	32,847	32,847	32,847	0
Progress Tracking (ProgTrk)	17,451	4,622	17,451	46,712	46,712	46,712	0
Procedure Supervision (ProcSup)	4,670	2,927	4,670	13,315	13,315	13,315	0
DynMan	36,472	7,042	72,944	72,944	74,203	54,708	18,236
Request Revision (ReqRev)	18,236	6,885	36,472	36,472	36,472	18,236	18,236
Interruption and Resumption (IntrRes)	18,236	6,692	36,472	36,472	37,731	36,472	0
SitInt	17,063	3,224	17,063	17,063	30,429	17,063	0
Affective Stance (AffSta)	6,603	2,876	6,603	6,603	10,620	6,603	0
Pragmatic Intent (PragInt)	5,880	1,973	5,880	5,880	10,241	5,880	0
Interaction Dynamics (IntDyn)	4,580	1,434	4,580	4,580	9,568	4,580	0
Total	159,998	11,162	224,161	281,551	371,901	263,315	18,236

The task families produce different kinds of interaction multiplicity. In **ConAmb**, 30,980 records contain 58,671 user turns, of which 56,344 belong to **AmbClar**. In **PerExe**, 35,484 single-turn records generate 92,874 required response annotations (2.62 per record), because a standing request can trigger multiple updates. In **DynMan**, every record contains two user turns, and all 18,236 forbidden response annotations occur in **ReqRev**, where a prior response obligation is superseded. Hence, turns and response annotations cannot be treated as interchangeable training units.

The 371,901 evidence citations comprise visual observations (33.6%), scene text/OCR (33.2%), speech (24.1%), and non-speech sound (9.0%), with percentages rounded independently. Relative

to the actual audible end of the spoken query, 31.1% of cited intervals end beforehand, 28.0% cross it, and 40.8% begin afterward. The aggregate hides task-specific timing patterns: 77.7% of **GrdRe** citations end before the actual audible query end, whereas all **PerExe** citations begin afterward. These interval-based proportions do not establish when a complete answer first becomes justified, and citation frequencies do not measure the independent importance of each modality. The distinct source videos have a median duration of 123.7 s (interquartile range: 83.7–172.5 s). Figure 5 visualizes the family-level composition and duration distribution.

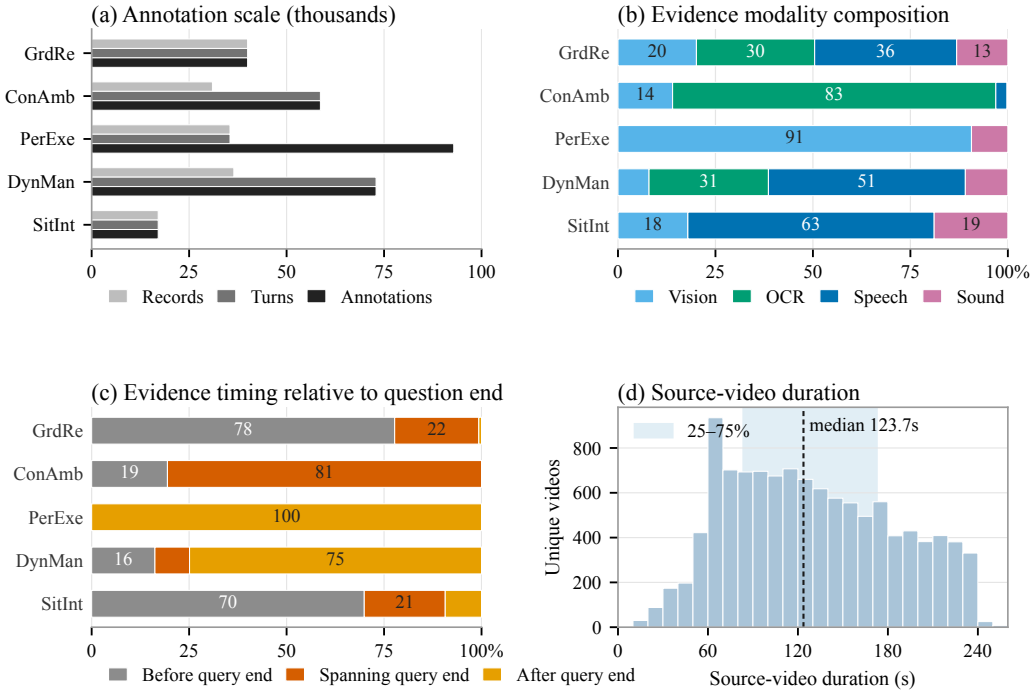


Figure 5: Overview of NativeQuery-Train. (a) Counts of dialogue records, spoken user-query turns, and response annotations by task family. (b) Modality of cited evidence and (c) location of cited evidence intervals relative to the actual audible query end, both normalized within each family. An interval crosses the actual audible query end if it starts before and ends after that boundary. (d) Duration of distinct source videos; the shaded band marks the interquartile range. Evidence panels count citations, including reuse across response annotations.

B Response-Triggering Analysis

Table 5 isolates response-triggering behavior. NativeStream-7B achieves the highest overall T_Prec./T_Acc./T_F1, with values of 30.1%, 57.8%, and 39.5%. MiniCPM-o-4.5-9B reaches 25.4%, 47.2%, and 32.8% on the same metrics. NativeStream-7B also leads all three metrics in every task family. The largest differences occur in DynMan and ConAmb. In DynMan, NativeStream-7B reaches 27.0%, 52.0%, and 35.5%, whereas MiniCPM-o-4.5-9B obtains 17.2%, 28.2%, and 21.4%. In ConAmb, the corresponding values are 40.5%, 70.1%, and 51.3% for NativeStream-7B and 34.5%, 56.5%, and 42.9% for MiniCPM-o-4.5-9B. These results indicate a stronger advantage when response decisions depend on evolving task state or conflicting and ambiguous evidence. The gap between T_Acc and T_Prec./T_F1 shows that reliable triggering remains difficult.

C Data Synthesis Prompts

We synthesize NativeQuery-Train with five task-specific prompt templates aligned with the benchmark taxonomy. Each template constrains the target capability, the information available at query time, the cross-modal evidence required, and the conditions under which no candidate should be pro-

Table 5: Comparison of response-triggering behavior on the NativeQuery-Bench. For each task family (GrdRe, ConAmb, PerExe, DynMan, and SitInt; Section 3.1) and the overall column, we report T_P., T_A., and T_F1. *Overall* is the macro average across the five task families. All values are percentages, and higher is better.

Model	GrdRe			ConAmb			PerExe			DynMan			SitInt			Overall		
	T_P. ↑	T_A. ↑	T_F1 ↑	T_P. ↑	T_A. ↑	T_F1 ↑	T_P. ↑	T_A. ↑	T_F1 ↑	T_P. ↑	T_A. ↑	T_F1 ↑	T_P. ↑	T_A. ↑	T_F1 ↑	T_P. ↑	T_A. ↑	T_F1 ↑
Proprietary Streaming Omni Models																		
Qwen 3.0-Omni Flash R	6.1	39.4	10.6	6.9	37.1	11.6	16.5	61.8	26.0	9.0	40.7	14.7	8.0	52.9	13.9	9.3	46.4	15.4
Qwen 3.5-Omni Plus R	6.5	35.6	11.0	7.3	33.8	12.0	17.8	63.1	27.8	9.7	36.7	15.3	8.6	54.0	14.8	10.0	44.6	16.2
Qwen 3.8-Omni Flash R	6.8	36.8	11.5	7.6	35.0	12.4	18.3	**66.5**	28.7	10.3	38.5	16.3	8.9	56.1	15.4	10.4	46.6	16.9
Gemini 2.5 Flash Live	8.1	42.8	13.6	10.9	38.6	17.0	21.8	38.9	27.9	11.3	36.2	17.2	9.7	47.6	16.1	12.4	40.8	18.4
Gemini 3.8 Live	8.7	40.5	14.3	11.8	35.3	17.6	23.2	35.4	28.0	12.0	34.0	17.7	10.2	50.2	17.0	13.2	39.1	18.9
Open-source Streaming Omni Models																		
ROMA-7B	7.6	44.7	13.0	8.4	42.1	14.0	19.8	64.9	30.3	11.1	40.3	17.4	8.8	**57.6**	15.3	11.1	49.9	18.0
MiniCPM-o-4.5-9B	30.2	67.5	41.7	34.5	56.5	42.9	27.5	44.8	34.0	17.2	28.2	21.4	17.5	39.0	24.2	25.4	47.2	32.8
NativeStream-7B	**31.0**	**68.0**	**42.6**	**40.5**	**70.1**	**51.3**	**31.0**	52.6	**39.0**	**27.0**	**52.0**	**35.5**	**21.0**	46.3	**28.9**	**30.1**	**57.8**	**39.5**

duced. Figure 6 specifies grounded multimodal reasoning; Figure 7 handles conflict, false premises, and ambiguity; Figure 8 defines persistent task execution; Figure 9 governs request revision, interruption, and resumption; and Figure 10 covers situated affective and pragmatic interpretation. Together, these prompts discourage single-modality shortcuts, future leakage, and unsupported inference while preserving natural single-turn QA outputs.

D Limitations and Future Work

D.1 Limitations

Our benchmark relies on synthesized spoken queries mixed into existing audio-visual streams, even though the query voices and timing are matched to the source context. It therefore only partially captures spontaneous human speech, overlapping conversation, and long-horizon interaction, while the current frame sampling and timing tolerance may miss fine-grained latency and interruption behavior.

D.2 Future Work

Future work will collect more natural human speech and extend the benchmark to longer, more diverse real-world interactions. We will also study lower-latency response triggering with finer-grained temporal evaluation and fully causal inference.

TASK A — Grounded Multimodal Reasoning

OBJECTIVE
Generate a natural conversational interaction focused on answering one specific user question. The answer must depend on complementary information from both original scene audio and visual input. Removing either modality must leave the answer unresolved or non-unique.

QUESTION SCOPE
The question may concern an event, state, object identity, temporal relation, count, comparison, reference grounding, or cross-temporal binding. Use natural spoken wording; pronouns, names, visual attributes, or sound references are allowed when they fit the context.

CAUSALITY
Set the query time t_q when the user starts speaking. The question may use target names, visual attributes, or context established at or before t_q, but must not reveal the final answer, a future identity, or a later outcome. It is valid for the user to ask about an unresolved event and for the assistant to answer after later evidence arrives.

EVIDENCE AND ANSWER
Use only the supplied evidence. Every cited observation must contribute to the answer; do not fill gaps with plausibility, narrative expectation, or temporal proximity. The answer must be concise, direct, and fully supported. If the evidence cannot resolve the question uniquely, or the question is answerable from one modality alone, return no candidate.

OUTPUT STYLE
Return one single-turn QA candidate in the shared QA format. Write `question` exactly as a natural user utterance, not as an exam question or annotation description. Write `answer` as one concise, direct, complete factual response. Do not include reasoning, evidence IDs, or extra dialogue turns in the spoken text.

Figure 6: Task A prompt for generating grounded multimodal questions and answers that require complementary audio-visual evidence and respect query-time causality.

TASK B — Conflict and Ambiguity Handling

OBJECTIVE
Generate a natural interaction in which the assistant handles incompatible information between modalities, corrects a grounded false premise, or resolves genuine ambiguity when multiple candidates fit the user's request.

CONFLICT
Treat audio and visual observations as conflicting only when they make incompatible claims about the same grounded entity, event, state, or result under the same time and scope. The answer should report what each modality supports and should not assume that audio or visual evidence is automatically authoritative. Correct the discrepancy only when the evidence establishes the relevant fact.

AMBIGUITY
When the combined audio and visual context leaves multiple reasonable candidates, the answer must be one short, targeted clarification question asking for the missing distinction. Both modalities must make a necessary contribution to establishing the request or its candidate interpretations.

FALSE PREMISE AND CAUSALITY
A false premise or misunderstanding must arise naturally from information available at t_q or from supplied dialogue context. Do not invent an arbitrary user mistake. The question must not leak future evidence, and the answer must be fully supported by the cited observations.

BOUNDARIES
Do not treat extraction noise, mismatched entities, different scopes, or ordinary uncertainty as a genuine conflict. Do not fabricate the user's reply or the resolved outcome. If no real conflict, grounded false premise, or legitimate ambiguity is supported, return no candidate.

OUTPUT STYLE
Return one single-turn QA candidate in the shared QA format. Write the user's `question` naturally and ground any mistaken premise in the available context. Write `answer` either as a concise evidence-based correction/report of the discrepancy or as one targeted clarification question. Do not fabricate a user reply, resolved outcome, chain-of-thought, or extra dialogue turns.

Figure 7: Task B prompt for handling cross-modal conflicts, false premises, and genuine ambiguity through evidence-based corrections or targeted clarification.

TASK C — Persistent Task Execution

OBJECTIVE
Generate a natural interaction in which the user establishes a standing request at t_q and the assistant continuously observes the stream. The assistant responds when a specified condition is met, the tracked state meaningfully changes, or an evidence-backed error occurs.

REQUEST TYPES
The request may involve:
1. Monitoring and notification: detect the first or next qualifying occurrence, a state change, or a cumulative condition.
2. Progress tracking: track a goal, checklist, or procedure as completed, pending, or unknown.
3. Guidance and correction: supervise an observable task, detect a deviation from the user's stated criteria, and provide an actionable correction.

GROUNDING AND CAUSALITY
The criteria must be established by the user's request, prior dialogue, or supplied evidence, not by an assumed external procedure, safety norm, or common sense. The question at t_q must not know whether or how future events will unfold. Do not leak future completion, future errors, or the eventual notification into the initial request.

CROSS-MODAL REQUIREMENT
The trigger, progress update, or error must require complementary audio and visual information. Removing either modality must leave the condition unresolved, the progress unverifiable, or the deviation undetectable. Do not treat one spoken claim, one visual state, or temporal proximity as sufficient when the criterion requires both.

OUTPUT BOUNDARY
Count distinct events correctly and do not duplicate a sustained state. Do not invent an event merely to force a response. If the condition never occurs, the correct behavior is continued observation; return no candidate for an untriggered request, unsupported criterion, or one-shot lookup.

OUTPUT STYLE
Return one single-turn QA candidate in the shared QA format. Write `question` as the user's natural standing request at t_q, without revealing whether the future condition will occur. Write `answer` as the concise notification, progress update, or actionable correction that becomes justified when the trigger occurs. Do not output a response for an untriggered condition, and do not include internal reasoning or extra dialogue turns.

Figure 8: Task C prompt for persistent monitoring, progress tracking, and task supervision, with responses triggered only by grounded state changes or errors.

TASK D — Dynamic Task Management

OBJECTIVE
Generate a conversational interaction in which the user modifies an established request through revision, cancellation, interruption, temporary task switching, or resumption. The assistant must update, preserve, or restore the relevant task state.

INTERACTION SCOPE
A revision may change the target, scope, or conditions of a prior request. A cancellation may replace the prior goal with a new target. An interruption may temporarily switch to another request, and a resumption must continue the paused goal using the state accumulated before the interruption.

PIVOT TURN
The user turn at t_q must operate on a prior request that is naturally implied by the wording. The question should represent the exact pivot, with one control statement followed by one actionable request. It must not use a media event as the target of cancellation, embed a description of the current media state merely to establish timing, or combine independent questions.

ANSWER REQUIREMENT
The answer must produce an immediate, substantive, media-grounded result under the revised, interrupted, or resumed conditions. A bare acknowledgment such as “Got it” is invalid. A resumed answer must reflect accumulated progress rather than restart from zero.

CROSS-MODAL AND CAUSALITY RULES
The underlying result must require complementary audio and visual evidence. Removing either modality must leave the result unresolved or incorrect. The question may use only information available up to t_q and must not leak future media state. Do not cite the user's implied prior request as evidence. Return no candidate when the timeline does not support a meaningful intent change with cross-modal consequences.

OUTPUT STYLE
Return one single-turn pivot QA candidate in the shared QA format. Write `question` as the user's natural revision, cancellation, interruption, switch, or resumption utterance, followed by one actionable request. Write `answer` as a concise substantive report under the new or resumed goal; never use a bare acknowledgment such as “Got it.” Do not include the implied prior request as a separate fabricated turn or expose internal reasoning.

Figure 9: Task D prompt for revising, cancelling, interrupting, switching, and resuming standing requests while preserving the relevant task state.

TASK E — Situated Interaction Reasoning

OBJECTIVE
Generate a natural interaction whose central information need is an interpretation of expressed or implied emotion, communicative attitude, non-literal meaning, or interpersonal dynamics. The answer must resolve social or affective meaning rather than merely report a factual observation.

TASK BOUNDARY
A question whose answer is only a speaker or addressee identity, literal speech content, visible action, scene object, event description, or future outcome is factual QA, not Task E. Such facts may support the interpretation but cannot be the requested result.

INTERPRETATION RULES
The answer may describe an observable emotion, attitude, interactional stance, response, or relational change, but must not claim access to hidden feelings, stable personality, true motives, or unobserved mental states. Limit the conclusion to the grounded interaction and use wording whose certainty matches the evidence.

CROSS-MODAL REQUIREMENT
The social or affective interpretation itself must require complementary audio and visual information from the same participant, interaction, and relevant temporal context. Audio and visual evidence must each contribute to the interpretation, not merely identify a person, event, or time.

If several interpretations remain reasonable, give a qualified conclusion. If the evidence does not support a reliable, genuinely cross-modal social interpretation, return no candidate.

OUTPUT STYLE
Return one single-turn QA candidate in the shared QA format. Write `question` as a natural request for social or affective interpretation. Write `answer` as a concise interpretation grounded in observable audio-visual cues, with qualified wording when the evidence is ambiguous. Do not merely repeat literal speech or visible actions, and do not state hidden motives or internal feelings as facts.

Figure 10: Task E prompt for interpreting grounded emotion, communicative attitude, and interpersonal dynamics from complementary audio-visual cues.