STREAMCAST: BENCHMARKING MLLMS FOR FUTURE PREDICTION IN INTERACTIVE LIVESTREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Livestreaming is a real-time and interactive form of video, where events evolve continuously through visual, audio, and audience signals. Understanding livestreams requires MLLMs to move beyond retrospective comprehension and anticipate near-future outcomes from evolving context, a capability underexplored by existing benchmarks. In this paper, we introduce **StreamCast**, a benchmark for evaluating near-future prediction in livestreams. Each video contains an anchor question targeting at the prediction of the immediate future, with support questions probing capabilities relevant to the prediction: *perception localization*, *interaction dynamics*, *historical understanding*, and *causal attribution*. Constructed through hierarchical multimodal representation, agentic task formation, and automatic verification, StreamCast comprises 470 real livestreams across six categories and 1,421 multiple-choice questions. We further assess benchmark quality through human evaluation of a sampled subset. We evaluate 15 open-source and proprietary MLLMs and analyze performance across diagnostic dimensions, modalities, and languages. The strongest model reaches 63.19% accuracy on future prediction in the Full-context setting while achieving 85.07–90.67% across the four diagnostic dimensions, highlighting the difficulty of anticipating future developments even when prerequisite capabilities are relatively strong. We release our code and dataset at `https://anonymous.4open.science/r/StreamCast`.

1 INTRODUCTION

Video understanding has advanced rapidly from recognizing actions and short events toward modeling the complex temporal dynamics of real-world video (An et al., 2026). Recent multimodal large language models (MLLMs) have further expanded video understanding by integrating visual, audio, and textual information, enabling richer interpretation of events and cross-modal relationships. These advances open new possibilities in real-time applications such as livestream content moderation (Wang et al., 2026a) and recommendation (Cao et al., 2026), where anticipating upcoming events can help identify potential risks and deliver timely, context-aware recommendations.

The evaluation of video understanding has traditionally focused on recognition, localization, and explanation of observed events in completed videos (Wu et al., 2024), as shown in Figure 1. Recent online video benchmarks have extended evaluation to evolving video streams, including timestamp-aware understanding and future status forecast (Li et al., 2025b). Livestreams represent a more open-ended real-time setting, where events continue to unfold as the stream progresses. Their evolution can further be shaped by ongoing interactions among streamers, audiences, and the surrounding context (Wang et al., 2025; 2026b). Anticipating subsequent developments therefore requires MLLMs to integrate multimodal evidence and reason over the evolving livestream context. Yet such interaction-aware future prediction remains insufficiently examined in existing benchmarks.

To address this gap, we build our benchmark around three key principles. 1) First, the benchmark should directly evaluate the near-future prediction capability of MLLMs: given a livestream up to the present moment, an MLLM should infer how the ongoing situation is likely to evolve, rather than merely recognizing isolated events that have already occurred. 2) Second, predictions should be grounded in evidence available before the prediction point. The answer should emerge from clues distributed across historical events, multimodal cues, and audience reactions, rather than relying on commonsense priors or open-ended speculations. 3) Third, the benchmark should provide diagnostic

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.



Figure 1: Existing benchmarks typically reasons about unfolded events, whereas StreamCast asks MLLMs to infer what will happen next from an ongoing livestream using only the preceding context.

insights beyond prediction accuracy. By decomposing prediction into underlying capabilities, it reveals whether a model fails due to missing key evidence, misunderstanding interactions, losing track of history event, or failing to attribute observed reactions to their causes.

Based on this formulation, we introduce **StreamCast**, a benchmark designed to evaluate the near-future prediction capability of MLLMs in livestream understanding. StreamCast organizes each video around an anchor question that targets the prediction of the immediate future trajectory of the livestream, while accompanying support questions that examine the prerequisite capabilities relevant to a successful prediction. These capabilities cover four key dimensions: perceiving relevant visual and acoustic evidence within the livestream video (*perception localization*), recognizing interactions among livestream participants (*interaction dynamics*), tracking the progression of earlier events over time (*historical understanding*), and attributing observed reactions to their underlying causes (*causal attribution*). By coupling prediction with diagnostic analysis, StreamCast enables identification of whether prediction failures primarily arise from insufficient perception of relevant multimodal evidence or from limitations in reasoning over the evolving livestream context.

To enable scalable and evidence-grounded construction, we develop an automated annotation pipeline that combines hierarchical multimodal representation, agentic task formation, and automatic verification. The resulting benchmark contains 470 real livestreams across six categories and 1,421 multiple choice questions (MCQ).We further conduct human validation on a sampled subset to ensure question quality and evidence support. Models are evaluated under two complementary settings: 1) Local setting, which provides a short time window that directly contains the evidence segment, isolating the model's ability to perform localized prediction. 2) Full-context setting, that provides the complete livestream as input, requiring tested models to retrieve relevant evidence and reason over the entire time span. Together, these settings disentangle failures in grounded prediction from failures in long-video evidence retrieval and reasoning over distributed temporal context.

Our contributions are summarized as follows:

- We introduce **StreamCast**, a benchmark for evaluating near-future prediction in interactive livestreams. By organizing each video around an anchor prediction question and diagnostic support questions, StreamCast evaluates models' ability to anticipate future developments in evolving livestreams while probing the prerequisite capabilities relevant to their predictions.
- We develop an evidence-grounded benchmark construction pipeline that combines multimodal representation, agentic task formation, and automatic verification into a process that enables scalable construction of prediction tasks grounded in evidence rather than commonsense speculation.
- We conduct extensive experiments on open-source and proprietary MLLMs, with fine-grained analyses across diagnostic dimensions and modalities. We further investigate whether improvements in prerequisite capabilities through fine-tuning can transfer to near-future prediction, providing insights into the relationship between these capabilities and prediction performance.

2 Related Work

Video Understanding Benchmark. Benchmarks for video understanding have progressed from short-clip question answering toward long and temporally demanding evaluation. NExT-QA (Xiao et al., 2021) targets causal and temporal question answering, MVBench (Li et al., 2023b) and Tem-

pCompass (Liu et al., 2024) probe fine-grained temporal skills, and EgoSchema (Mangalam et al., 2023), Video-MME (Fu et al., 2024a), MMBench-Video (Fang et al., 2024), and LongVideoBench (Wu et al., 2024) extend evaluation to long-form video. A recent line begins to consider streaming or online settings in which frames arrive sequentially. VideoLLM-online (Chen et al., 2024) formulates online dialogue over streaming video; StreamingBench (Lin et al., 2024) and OVO-Bench (Li et al., 2025b)extend evaluation to online video settings, where OVO-Bench further includes future prediction based on the current visual state; SVBench (Yang et al., 2025) targets long-horizon memory through multi-turn dialogue; OmniMMI (Wang et al., 2025) and OVBench (Huang et al., 2025) further probe proactive abilities such as real-time narration and alerting. Recently, LiViBench (Wang et al., 2026b) brings audience interactions into livestream evaluation, offering user comment as auxiliary information. Distinct from of these benchmarks, we focus on the evaluation of whether a model can infer upcoming developments from the preceding multimodal context, and further provide diagnostic insights into the prerequisite capabilities underlying such prediction.

Multimodal Large Language Models. Multimodal large language models (MLLMs) extend language models with visual perception by aligning a visual encoder to a language backbone, as in BLIP-2 (Li et al., 2023a) and LLaVA (Liu et al., 2023), and subsequently scale to stronger encoders and larger instruction corpora in Qwen-VL (Bai et al., 2023), Qwen2.5-VL (Bai et al., 2025b), InternVL3 (Zhu et al., 2025), and Gemini (Gemini Team, 2023). To process temporal input, a line of video MLLMs extends this recipe from images to frame sequences: Video-LLaVA (Lin et al., 2023) unifies image and video representation, VideoLLaMA3 (Zhang et al., 2025) and LLaVA-Video (Zhang et al., 2024b) strengthen video instruction tuning, and LongVA (Zhang et al., 2024a) transfers long-context ability from text to many-frame video. A further line moves toward omni-modal understanding that jointly consumes vision, audio, and text. Early efforts such as Qwen2-Audio (Chu et al., 2024) and video-SALMONN (Sun et al., 2024) add speech and audio understanding, while VITA (Fu et al., 2024b) and Baichuan-Omni (Li et al., 2024) integrate the modalities within a single interactive model. More recent omni-modal models further close the gap to proprietary systems, including Ola (Liu et al., 2025b), Baichuan-Omni-1.5 (Li et al., 2025a), Ming-Omni (Inclusion AI, 2025), and the Qwen2.5-Omni (Xu et al., 2025a) and Qwen3-Omni (Xu et al., 2025b) series. These models broaden perception and long-context reasoning, yet they are trained and evaluated almost entirely on retrospective understanding of already-observed content, whether their audio-visual grounding supports inferring near-future developments from an ongoing and interactive stream remains untested, exactly the capability we aim to measure.

3 METHODOLOGY

3.1 OVERVIEW

Figure 2 illustrates the dataset construction procedure of StreamCast, which consists of three stages: (i) data collection of real-world livestreams spanning diverse categories and content dynamics 3.2; (ii) hierarchical multimodal representation, which converts each long, raw stream into a time-aligned textual annotation of visual semantics, speech, and audience actions, together with statistical interaction signals that highlight information-dense moments 3.3; and (iii) agentic task formation, in which an agent dynamically retrieves the context needed for each question, generates evidence-grounded anchor and support questions, and automatically validates their quality 3.4.

Problem Statement. StreamCast evaluates whether MLLMs can predict the near-future evolution of a livestream from its preceding multimodal context. We formalize a livestream as a temporally ordered multimodal signal $L = (f, a, i)$, where f denotes the stream of visual frames, a the audio stream carrying speech and ambient sound, and i the stream of audience interactions such as comments, gifts, and likes. Given a prediction point t, an anchor question q_{anchor} asks how the ongoing situation will evolve in the immediately following interval Δ, while each support question q_s probes a prerequisite capability drawn from four dimensions: *perception localization*, *interaction dynamics*, *historical understanding*, and *causal attribution*. Each question is defined with respect to the preceding context $L_{<t}$, which contains all multimodal evidence available for answering it. For each question q, we require the available context to support a unique target answer within the corresponding candidate set:

$$y^* = g(q, L_{<t}, \mathcal{Y}_q),$$

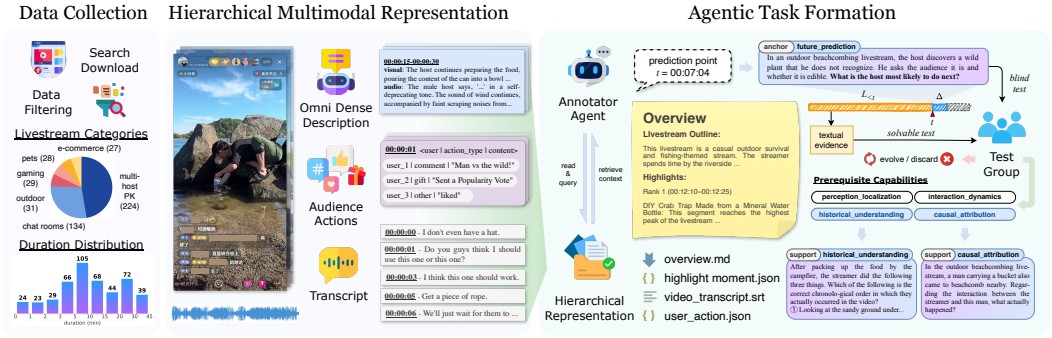


Figure 2: Overview of StreamCast construction pipeline. Raw livestreams are filtered and transformed into hierarchical multimodal representations, from which an annotator agent read and retrieve context to set a suitable prediction point. Then the agent constructs anchor & support question and conduct blind and solvable tests ensure evidence-grounded and answerable questions.

where $\mathcal{Y}_q$ denotes the denotes a question-conditioned candidate answer space of q, and g represents the task-specific decision rule induced by the question formulation and its available evidence. For anchor questions, this target corresponds to the near-future outcome that the preceding interaction state most strongly supports, as $\mathcal{Y}_q$ constraints the future to a finite set of plausible continuations. No signal after t is exposed to the tested model or treated as input evidence, ensuring that StreamCast evaluates evidence-grounded prediction rather than open-ended speculation.

3.2 Data Collection

To construct a realistic benchmark for evolving livestream scenarios, we collect real-world livestream videos from Bilibili[1] covering diverse categories and interaction patterns. We restrict our collection to publicly accessible archived recordings that have remained consistently available on the platform, ensuring stable data provenance and reproducibility of the benchmark. We further clean and filter the recordings to remove edited or post-produced videos and retain only raw livestream recordings that preserve the original temporal flow and real-time interactions. The resulting corpus contains 470 livestreams spanning six categories: multi-host PK, chat rooms, outdoor, gaming, pets, and e-commerce, which differ substantially in pacing, dominant modality, and interaction style.

3.3 Hierarchical Multimodal Representation

Constructing evidence-grounded future prediction questions requires linking future outcomes to the preceding evidence that supports their prediction. In particular, question q should be grounded in concrete multimodal evidence, while candidate answer space $\mathcal{Y}_q$ should be plausible and distinguishable through the preceding context, so that models must reason over the actual livestream rather than rely on commonsense priors or superficial guessing. This is especially challenging for livestreams, where relevant evidence may be distributed across multiple modalities, including visual events, speech, audio cues, and audience interactions across different points in time. We therefore systematically convert these heterogeneous multimodal signals into a time-aligned hierarchical multimodal representation in textual form, which preserves the rich semantic information as well as their temporal relationships to facilitate subsequent task formation and validation.

As shown in Figure 2,we construct this representation from three complementary sources of information extracted from the visual, audio, and interaction streams f, a, and i. We first divide each livestream into short chunks and generate dense omni descriptions of the video content, representing the aligned visual and audio at a segment level. These segments are annotated sequentially to capture fine-grained event progression, visual details, and non-speech audio cues. Next, we extract frame-level audience interactions, including comments, gifts, and likes, and represent each interaction as a structured entry $\langle \text{user}, \text{action type}, \text{content} \rangle$. Duplicate interactions across temporally adjacent frames are removed through fuzzy matching. We further obtain speech transcripts using automatic speech recognition (ASR) to capture spoken content in greater detail. These three sources are then

[1] https://www.bilibili.com/

temporally aligned and integrated into a unified representation in which each segment contains its visual and audio description, corresponding speech, and audience interactions.

We further derive temporal interaction statistics from the audience track to identify information-dense moments for subsequent task construction. Since prediction and diagnostic questions are most informative when centered on consequential developments rather than arbitrary timestamps, we use changes in audience activity as an initial signal for locating candidate moments. Specifically, we measure the interaction rate over time and detect local bursts using a z-score criterion. We further apply an LLM-based semantic filter to retain moments associated with salient developments in the stream. The resulting highlights serve as candidate points for task formation, allowing the annotator agent to focus on temporally localized contexts where event progression, audience responses, and multimodal evidence are most informative. Together, the multimodal annotations and interaction highlights provide a hierarchical view of the livestream, spanning fine-grained contextual evidence and overall event progression. This representation supports subsequent task formation by allowing the annotator agent to examine both local details and broader temporal context. We provide annotation details and specific model settings used during this process in Appendix B.

3.4 Agentic Task Formation

Based on the hierarchical representation, we develop an agentic task formation pipeline to construct evidence-grounded test sets and automatically verify their quality. The annotator agent first examines the entire livestream and identifies a suitable prediction point t, corresponding to a meaningful transition in the ongoing situation where a salient development is about to unfold. This point defines the moment at which prediction is made: the preceding context $L_{<t}$ represents what has been observed so far, while the events after t constitute the future to be anticipated. Based on this point, the agent constructs an anchor question, the primary question in each question set, which asks the model to predict how the ongoing situation will develop in the following interval Δ. Along with the question, the annotator agent is also asked to provide the evidence e_{anchor}, a set of textual context for answering the anchor, where e_{anchor} is restricted to $L_{<t}$, thereby strictly preserving the temporal asymmetry between observed context and future outcome.

To ensure the prediction is dependent on the livestream context, the anchor question are validated if all candidate answers describe alternative outcomes within the same decision space $\mathcal{Y}_q$. That is, the options correspond to different plausible developments of the same ongoing situation, rather than unrelated events that can be distinguished from their surface semantics. For each q_{anchor}, $L_{<t}$ should therefore include clues for one possible outcome while ruling out the other alternatives. We enforce these properties through two complementary automatic tests using text-only models with independent test groups to reduce model-specific bias. In the *blind test*, the test group receives only the question and candidate answers without any livestream evidence, and the question passes if the group cannot reliably determine the answer, ensuring that the question cannot be solved from the options or general knowledge alone. In the *solvable test*, the group is additionally provided with textual evidence e_{anchor} and must identify the intended answer, guaranteeing the question can be solved by the information in $L_{<t}$. Together, these tests ensure that the anchor is both dependent on the preceding context and sufficiently constrained by the evidence available at the prediction point.

The validated future prediction question q_{anchor} and evidence e_{anchor} then serves as the basis for diagnostic task formation. The agent analyzes the evidence and reasoning involved in the anchor prediction, drawing on other relevant portions of $L_{<t}$ when needed, to construct $n \in 2, 3, 4$ support questions $q_s^1, ..., q_s^n$. Each support question targets one of four diagnostic dimensions and is paired with an evidence-supported correct answer and plausible distractors. The support questions undergo the same blind and solvable tests as the anchor, allowing their results to provide diagnostic evidence about the perceptual and reasoning capabilities relevant to prediction meanwhile ensuring the quality of these tasks. The detailed algorithm and diagnostic capability definition is provided in Appendix B.

4 Experiment

4.1 Experimental Settings

We evaluate 15 representative omni-modal models, including 5 proprietary and 10 open-source models. Each model is evaluated under two settings that pose identical questions but differ in the context

Table 1: Main results on StreamCast under the *Local* (*Loc.*) and *Full* context settings, reported as accuracy (%). † Parameters are not officially disclosed, estimated values are provided.

	Causal Attribution		Historical Understanding		Interaction Dynamics		Perception Localization		Future Prediction	
	Loc.	*Full*	*Loc.*	*Full*	*Loc.*	*Full*	*Loc.*	*Full*	*Loc.*	*Full*
Proprietary Models										
Gemini-3.1-Pro-Preview	85.59	**87.29**	85.00	**90.67**	**84.16**	**85.07**	84.29	**88.14**	**62.77**	**63.19**
Gemini-2.5-Flash (Gemini Team, 2025)	83.90	77.12	82.00	75.00	75.57	76.47	76.60	75.64	46.17	44.26
Qwen3.5-Omni-Plus (Team, 2026)	**89.83**	**87.29**	**87.33**	87.67	83.71	83.26	83.65	85.58	59.15	55.53
Qwen3.5-Omni-Flash (Team, 2026)	83.90	28.81	79.00	29.67	76.92	25.79	82.37	27.56	45.32	14.04
Doubao-Seed-2.0-Lite	68.64	77.12	66.00	69.67	79.64	74.66	68.91	71.15	59.57	59.36
Open-Source Models										
MiMo-V2.5 (310B/15B) (Xiaomi MiMo Team, 2026)	83.05	82.20	80.33	77.67	80.09	80.09	**84.94**	83.65	60.85	58.51
Qwen2.5-Omni (7B) (Xu et al., 2025a)	77.12	59.32	68.33	58.67	68.78	54.30	67.95	50.00	31.91	31.49
MiniCPM-o-4.5 (9B) (Cui et al., 2026)	75.42	50.00	70.33	56.00	67.42	51.13	70.83	54.81	31.70	31.91
Baichuan-Omni-1.5 (11B) (Li et al., 2025a)	64.41	52.54	62.67	52.00	62.90	47.96	66.35	43.27	28.09	28.72
OmniVinci (9B†) (Ye et al., 2025)	67.80	57.63	64.33	57.00	59.73	52.04	55.45	50.64	40.43	40.21
Ola (7B) (Liu et al., 2025b)	64.41	49.15	58.00	47.00	52.49	43.44	47.76	34.94	24.68	24.04
Nexus-O (7B†) (Liu et al., 2025a)	63.56	46.61	53.00	46.00	63.80	49.32	65.71	43.91	32.55	31.70
HumanOmni (7B) (Zhao et al., 2025)	53.39	41.53	42.00	38.33	39.37	28.05	48.40	32.69	26.60	25.11
Phi-4-Multimodal (5.6B) (Microsoft, 2025)	52.54	38.98	58.33	40.33	52.94	42.08	55.13	34.29	22.98	21.28
VITA-1.5 (7B) (Fu et al., 2025)	07.63	00.00	03.33	00.00	10.41	00.45	18.91	00.00	01.91	00.00

provided. In the Local setting, the model is given a short window that only contains the necessary evidence relevant to the question, so its accuracy reflects grounded prediction. In the Full setting, the model is given the entire preceding livestream $L_{<t}$ and must locate the relevant evidence itself before answering; the Local-to-Full difference therefore isolates a model's evidence-localization ability. We report accuracy on each of the five capability dimensions using English question and answer pairs by default. Detailed experimental settings are provided in Appendix A.

4.2 Main Results

From the main results in Table 1, we observe that: 1) Models exhibit substantial variation in livestream understanding, with future prediction remaining challenging. Gemini-3.1-Pro-Preview achieves the highest future prediction accuracy of 63.19% under the Full setting, while MiMo-V2.5 leads open-source models at 58.51%, surpassing several proprietary models. Despite strong performance on the diagnostic dimensions, substantial room for improvement remains in anticipating future developments. 2) Longer context disproportionately affects smaller models, while stronger models remain relatively stable. For instance, Phi-4-Multimodal drops from 55.13% to 34.29% on perception localization, whereas Gemini-3.1-Pro-Preview improves from 84.29% to 88.14%. Interestingly, future prediction accuracy changes relatively little for most models, possibly because prediction-relevant evidence is often concentrated near the prediction point. The larger declines on diagnostic dimensions suggest that accessing and integrating evidence across longer livestreams remains challenging, particularly for smaller models. 3) Model performance on the diagnostic dimensions is broadly consistent with their future prediction performance, but prediction remains systematically more difficult. Even Gemini-3.1-Pro-Preview achieves 85.07–90.67% across the diagnostic dimensions while reaching only 63% on future prediction. This gap suggests that the capabilities assessed by the support questions provide a necessary foundation for prediction, whereas anticipating the future further requires extrapolating beyond the observed interaction state. We also observe several anomalous results in VITA-1.5 and Qwen3.5-Omni-Flash, which arise from model output failures under specific settings. VITA-1.5 produces degenerate and repetitive tokens under the Local setting, while under the Full setting it often generates free-form scene descriptions without an option letter, resulting in responses that cannot be parsed as valid answers. Qwen3.5-Omni-Flash, in contrast, exhibits a substantial gap between the Local and Full settings because it frequently returns empty outputs when processing long videos. We report these results as observed without post-hoc filtering, thereby faithfully reflecting the models' actual performance under our evaluation protocol.

4.3 Modality Ablation

To understand how different modalities contribute to livestream understanding, we conduct modality ablation studies with both vision-language and audio-language models, examining the respective roles of visual and acoustic information in reasoning about evolving livestream events. We con-

Table 2: Modality ablation accuracy (%) under Local setting of vision-language and audio-language models, which isolate the visual and acoustic channels respectively.

	Causal Attribution	Historical Understanding	Interaction Dynamics	Perception Localization	Future Prediction
Visual Language Models					
GPT-5	**55.93**	43.67	54.75	41.67	**38.94**
Qwen3-VL-Instruct (8B) (Bai et al., 2025a)	50.85	**49.67**	**66.52**	**53.21**	29.79
Qwen2.5-VL-Instruct (7B) (Bai et al., 2025b)	42.37	45.67	59.28	50.96	27.87
InternVL3 (8B) (Zhu et al., 2025)	51.69	49.33	60.18	53.21	34.47
VideoLLaMA3 (7B) (Zhang et al., 2025)	45.76	43.00	51.13	43.91	35.32
LLaVA-Video-Qwen2 (7B) (Zhang et al., 2024b)	42.37	41.67	46.15	44.23	26.17
MiniCPM-V-2.6 (8B) (Yao et al., 2024)	45.76	40.33	52.49	40.06	22.98
LongVA (7B) (Zhang et al., 2024a)	38.14	33.00	31.67	28.21	22.77
Audio Language Models					
GPT-Audio	**79.66**	**71.00**	**56.56**	**58.65**	**40.00**
GPT-Audio-Mini	64.41	54.33	47.51	46.79	24.26
Qwen2-Audio-Instruct (7B) (Chu et al., 2024)	43.22	37.33	39.82	44.23	22.77
SALMONN (7B) (Tang et al., 2023)	41.53	37.67	35.75	38.78	24.26
Step-Audio-2-mini (8B) (Wu et al., 2025)	35.59	34.33	36.20	26.60	27.45
LTU-AS (7B) (Gong et al., 2023)	33.90	29.00	26.70	25.64	24.47

duct these comparisons under the Local setting to control for differences in long-context processing and evidence localization, thereby isolating the contribution of each modality given access to the same relevant context. From Table 2, we observe that: 1) Acoustic information provides a strong signal for livestream understanding. Unlike many conventional video understanding tasks that are primarily driven by visual content, the audio-only GPT-Audio achieves 56.5% overall accuracy and 40.0% on future prediction, outperforming the vision-only GPT-5 in both settings. This result suggests that speech and other acoustic cues play a central role in conveying how livestream events and interactions evolve over time. 2) Visual and acoustic information provide complementary evidence, while neither modality alone is sufficient for reliable future prediction. Within the Qwen family, Qwen2.5-VL performs substantially better than Qwen2-Audio on interaction dynamics, whereas the omni-modal Qwen2.5-Omni achieves markedly stronger overall performance than either single-modality variant. However, future prediction remains below 40% for all modality ablations, indicating that anticipating future developments requires jointly integrating complementary multimodal evidence rather than relying on a single information source.

4.4 Language Alignment Analysis

Livestream understanding relies heavily on language-dependent signals, particularly in speech and audience interactions, which provide important context for interpreting ongoing events. StreamCast livestreams are collected from Chinese platform, whereas the main experiments use English prompts for a unified evaluation interface across models, which potentially introducing a cross-lingual gap. We therefore compare performance under Chinese and English prompt settings.

Table 3 presents model performance under Chinese prompt setting and the gap versus English setting (*Loc.* in Table 1) , from the results we make three observations: 1) Prompt language introduces a measurable but moderate performance gap. Evaluating the same Chinese livestream content with English changes overall accuracy by 1.2% on average and by up to approximately 4% across models, while the difference can reach 8.5% on an individual dimension. These results confirm that language gap can affect livestream understanding, although its impact on overall performance remains bounded. 2) The gap varies across task dimensions and model families. The largest variations occur on the content-intensive Perception and causal attribution dimensions, for example, OmniVinci drops by 8.5% on causal attribution, whereas interaction dynamics and historical understanding are comparatively stable. Models developed outside the Chinese language ecosystem, including Gemini-3.1-Pro-Preview, Gemini-2.5-Flash, OmniVinci, and Phi-4-Multimodal, also consistently perform at least as well under Chinese prompts. This pattern suggests that querying Chinese livestream content in English can introduce additional cross-lingual uncertainty, particularly when understanding depends on fine-grained linguistic content. 3) Despite these differences, prompt language has little effect on the overall model ranking. The two settings produce nearly identical model rankings (Spearman $\rho = 0.98$), with top three remaining unchanged and only minor local swaps

Table 3: Language alignment accuracy (%) under Chinese-language prompts in Local setting. Subscripts give the difference from the English-prompt Local results in Table 1 (en - zh), with positive value indicates higher performance under English prompt setting.

	Causal Attribution	Historical Understanding	Interaction Dynamics	Perception Localization	Future Prediction
Proprietary Models					
Gemini-3.1-Pro-Preview	**87.29**$_{-1.70}$	86.33$_{-1.33}$	85.07$_{-0.91}$	83.97$_{+0.32}$	**65.96**$_{-3.19}$
Gemini-2.5-Flash (Gemini Team, 2025)	82.20$_{+1.70}$	83.33$_{-1.33}$	78.73$_{-3.16}$	78.85$_{-2.25}$	45.96$_{+0.21}$
Qwen3.5-Omni-Plus (Team, 2026)	**87.29**$_{+2.54}$	**88.00**$_{-0.67}$	**86.88**$_{-3.17}$	**85.26**$_{-1.61}$	60.43$_{-1.28}$
Qwen3.5-Omni-Flash (Team, 2026)	82.20$_{+1.70}$	79.33$_{-0.33}$	74.66$_{+2.26}$	79.81$_{+2.56}$	35.74$_{+9.58}$
Doubao-Seed-2.0-Lite	77.12$_{-8.48}$	67.33$_{-1.33}$	80.09$_{-0.45}$	69.55$_{-0.64}$	62.77$_{-3.20}$
Open-Source Models					
MiMo-V2.5 (310B/15B) (Xiaomi MiMo Team, 2026)	81.36$_{+1.69}$	84.00$_{-3.67}$	84.62$_{-4.53}$	84.29$_{+0.65}$	61.06$_{-0.21}$
Qwen2.5-Omni (7B) (Xu et al., 2025a)	78.81$_{-1.69}$	66.33$_{+2.00}$	67.42$_{+1.36}$	72.76$_{-4.81}$	29.15$_{+2.76}$
MiniCPM-o-4.5 (9B) (Cui et al., 2026)	77.12$_{-1.70}$	72.67$_{-2.34}$	67.87$_{-0.45}$	72.44$_{-1.61}$	27.45$_{+4.25}$
Baichuan-Omni-1.5 (11B) (Li et al., 2025a)	64.41$_{+0.00}$	61.00$_{+1.67}$	60.18$_{+2.72}$	63.78$_{+2.57}$	31.49$_{-3.40}$
OmniVinci (9B†) (Ye et al., 2025)	76.27$_{-8.47}$	69.00$_{-4.67}$	65.16$_{-5.43}$	61.22$_{-5.77}$	40.43$_{+0.00}$
Ola (7B) (Liu et al., 2025b)	63.56$_{+0.85}$	60.33$_{-2.33}$	52.49$_{+0.00}$	50.64$_{-2.88}$	23.19$_{+1.49}$
Nexus-O (7B†) (Liu et al., 2025a)	69.49$_{-5.93}$	51.33$_{+1.67}$	62.44$_{+1.36}$	68.27$_{-2.56}$	30.85$_{+1.70}$
HumanOmni (7B) (Zhao et al., 2025)	54.24$_{-0.85}$	44.67$_{-2.67}$	38.46$_{+0.91}$	50.96$_{-2.56}$	25.96$_{+0.64}$
Phi-4-Multimodal (5.6B) (Microsoft, 2025)	52.54$_{+0.00}$	56.00$_{+2.33}$	53.39$_{-0.45}$	53.21$_{+1.92}$	25.74$_{-2.76}$
VITA-1.5 (7B) (Fu et al., 2025)	07.63$_{+0.00}$	03.67$_{-0.34}$	09.95$_{+0.46}$	18.27$_{+0.64}$	01.91$_{+0.00}$

Table 4: Human assessment of StreamCast questions. For each criterion we report the mean score on a 1–3 scale and the inter-annotator consistency (pairwise exact-agreement rate).

Criterion	*Relevance*	*Evidence Support*	*Unambiguity*	*Difficulty*	*Modality Variety*	*Supportiveness*
Mean	2.89	2.76	2.71	2.55	2.90	2.72
Consistency (%)	88.4	70.0	64.7	44.0	83.3	53.3

among the more language-sensitive models. Therefore, while English prompting introduces a measurable cross-lingual gap, the main benchmark comparisons remain robust to prompt language.

5 Benchmark Analysis

5.1 Quality Analysis

Beyond automatic verification, we conduct an additional human assessment to further evaluate the quality of StreamCast data and examine whether the generated questions are faithful to the livestream, well-formed, or whether the anchor-support structure provides meaningful diagnostic information. We sample 20 livestreams spanning six categories, comprising 20 anchor and 49 support questions, and apply 5 expert annotators proficient in both Chinese and English. Annotators first answer each question based on the livestream video ($L_{<t}$), matching the information accessible to evaluated models; they then inspect the full livestream and intermediate textual annotations to assess question grounding and quality on a three-point scale (1 for the worst and 3 for the best). We assess six criteria covering content quality and benchmark structure: *relevance*, *evidence support*, *unambiguity*, *difficulty*, *modality variety*, and *supportiveness*, with rating criteria detailed in Table 8. We also provide additional human language fidelity assessment in Appendix E.

As shown in Table 4, the generated questions receive high ratings for relevance, evidence support, and unambiguity, indicating that they are generally faithful to the livestream content and admit well-defined answers. Modality variety and supportiveness also receive strong ratings of 2.90 and 2.72, respectively, suggesting that the questions draw on complementary information sources and that the support questions remain meaningfully connected to their corresponding anchors. The independent human answers provide an additional reference for task difficulty. Annotators correctly answer 82.7% of the support questions but only 40.0% of the anchors, consistent with the anchor–support gap observed in model evaluations. Difficulty also receives the lowest mean score and the lowest inter-annotator agreement, reflecting variation in how annotators perceive the difficulty of individual questions. Supportiveness likewise shows relatively lower agreement despite its high mean rating,

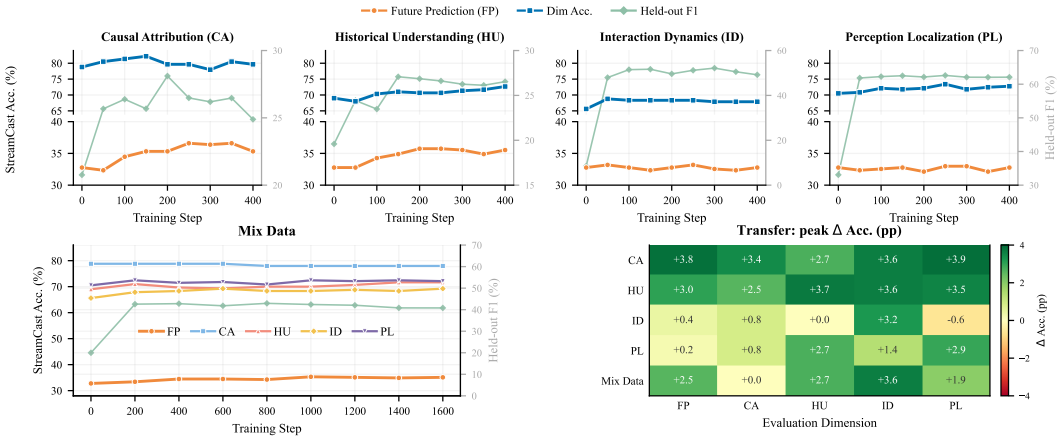


Figure 3: Capability transfer from targeted post-training to StreamCast. Held-out F1 tracks learning on each auxiliary source task; Dim Acc. and FP measures transfer to the corresponding StreamCast diagnostic dimension and anchor prediction. Mix Data setting jointly trains on all four supervision sources. The heatmap shows peak accuracy changes from the base model across checkpoints (pp).

suggesting that annotators generally recognize the support–anchor connection but differ in judging its strength. Nevertheless, these disagreements are typically small, rarely exceeding one rating level.

5.2 Capability Transfer

As the diagnostic dimensions in StreamCast are derived from the evidence and reasoning prerequisites of anchor predictions, we further investigate whether training on related capabilities can transfer to near-future prediction performance. With Qwen2.5-Omni-7B as the base model, we fine-tune separate variants on auxiliary public datasets selected to provide supervision relevant to each diagnostic dimension. We monitor held-out F1 to assess training progress and evaluate the resulting models on both support questions and anchor predictions in StreamCast. We additionally examine joint training across all four dimensions, with implementation details provided in Appendix C.

Figure 3 shows that causal attribution and historical understanding training improves the corresponding diagnostic dimensions and yields the largest peak gains on future prediction by +3.8% and +3.0%. In contrast, interaction dynamics and perception localization show clear learning progress on their auxiliary held-out sets but limited gains on the corresponding StreamCast dimensions, with similarly small prediction improvements of +0.4% and +0.2%. This pattern suggests that prediction gains broadly track whether training benefits transfer to the diagnostic dimensions themselves, rather than merely whether the auxiliary task is learned. Joint training also improves prediction but does not surpass causal attribution or historical understanding training alone. Overall, these results provide preliminary empirical support for the anchor–support decomposition as a diagnostic view of capability transfer to future prediction. Addtional analysis is provided in Appendix C.

6 Conclusion

We introduced StreamCast, a benchmark for evaluating interactive near-future prediction in livestreams. StreamCast organize each livestream around an anchor question that requires predicting how the ongoing situation will evolve, together with diagnostic support questions that assess the perception and reasoning capabilities underlying successful prediction. The dataset is constructed through hierarchical multimodal representation, agentic task formation, and automatic verification, resulting in 1,421 evidence-grounded questions over 470 real livestreams across six categories. We evaluate 15 state-of-the-art omni-modal MLLMs, covering both proprietary and open-source models, and use the anchor-support structure to provide fine-grained analyses of their strengths and limitations in livestream understanding. We further conduct modality analysis to examine how visual and acoustic information contribute to understanding and predicting evolving livestream events, and also language-alignment analysis to assess the potential impact of querying Chinese livestream content with English prompts. Finally, we further conduct human evaluation that verifies the quality of the automatically constructed questions and the reliability of the benchmark.

AI USE STATEMENT

We used generative AI tools in several parts of this work. LLMs were used to clean and filter data and, as described in the main paper and appendix, Claude Code was used as an annotator to construct benchmark questions, and assisted with software implementation. We used ChatGPT to polish the writing and check grammar. Generative AI tools were additionally used for literature retrieval, translation, and checking paper formatting. They were not used to formulate the research idea, design the methodology or experiments, or analyze and interpret the experimental results.

All AI-assisted outputs were reviewed by the authors. Benchmark data underwent the sampled quality review and validation as reported in the paper. AI-assisted code was executed and validated through the experiments, and the authors reviewed the polished text paragraph by paragraph and checked its factual claims and citations. Retrieved literature and bibliographic information were checked against the original publications, translations were reviewed by the authors, and experimental figures were cross-checked against results. The authors take responsibility for the final content of this work, including all text, code, data, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

StreamCast is constructed from livestream recordings uploaded by third parties and publicly accessible on the original platform. To limit redistribution and privacy risks, we do not redistribute the videos or release automated data acquisition scripts. We release only the annotations needed for evaluation, the corresponding timestamps, and links to the publicly available source videos. The released annotations exclude usernames, profile images, verbatim audience comments, streamer identities, and other identifying information. Rights to the source videos remain with their respective creators and the platform, and access to each source remains subject to the platform's terms and the continued availability of the original upload. Detailed documentation of data provenance, released and withheld artifacts, privacy considerations, licensing scope, and takedown procedures is provided in the Data Card included in the repository.

Because livestreams are created in uncontrolled real-world settings, some linked source videos may contain offensive, sensitive, or otherwise inappropriate material. We do not claim that such content has been exhaustively filtered and will document this possibility in the dataset materials so that users can make informed decisions about access and use. All sampling-based review and human evaluation reported in this work were conducted by the paper's authors; no external annotators or study participants were recruited.

REPRODUCIBILITY STATEMENT

We provide the complete evaluation questions and temporal annotations, the benchmark construction and evaluation workflow, model configurations, evaluation code, dataset statistics, and links to the publicly accessible source videos. The main paper describes the task formulation, construction pipeline, and experimental protocol, while the appendix provides the detailed annotation, verification, and evaluation procedures. Together, these materials support reproduction of the benchmark evaluation and the reported analyses. We do not release the automated data acquisition and processing code because doing so could create platform and legal risks. Consequently, reconstruction of the raw corpus also depends on the continued availability of third-party source videos; this limitation does not affect use of the released questions, annotations, evaluation code, and documented workflow on available sources. The anonymous repository includes the benchmark questions, source metadata, preprocessing and evaluation code, and a Data Card documenting the scope and limitations of the release.

REFERENCES

Zhaochong An, Zirui Li, Mingqiao Ye, Feng Qiao, Jiaang Li, Zongwei Wu, Vishal Thengane, Chengzu Li, Lei Li, Luc Van Gool, Guolei Sun, and Serge Belongie. Video understanding: From geometry and semantics to unified models. *arXiv preprint arXiv:2603.17840*, 2026.

Jinze Bai et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jiangxia Cao, Ruochen Yang, Xiang Chen, Changxin Lao, Yueyang Liu, Yusheng Huang, Yuanhao Tian, Xiangyu Wu, Shuang Yang, Zhaojie Liu, et al. Foresight prediction enhanced live-streaming recommendation. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pp. 1078–1082, 2026.

Joya Chen et al. Videollm-online: Online video large language model for streaming video. *arXiv preprint arXiv:2406.11816*, 2024.

Yunfei Chu et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Xinyu Fang et al. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024a.

Chaoyou Fu et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024b.

Chaoyou Fu et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yuan Gong et al. Joint audio and speech understanding. *arXiv preprint arXiv:2309.14405*, 2023.

Zhenpeng Huang et al. Online video understanding: Ovbench and videochat-online. *arXiv preprint arXiv:2501.00584*, 2025.

Inclusion AI. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Yue Jiang, Jichu Li, Yang Liu, Dingkang Yang, Feng Zhou, and Quyu Kong. Danmakutppbench: A multi-modal benchmark for temporal point process modeling and understanding. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. arXiv:2505.18411.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017.

Junnan Li et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023b.

Yadong Li et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024.

Yadong Li et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025a.

Yifei Li et al. Ovo-bench: How far is your video-llms from real-world online video understanding? *arXiv preprint arXiv:2501.05510*, 2025b.

Bin Lin et al. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Junming Lin et al. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

Che Liu et al. Nexus: An omni-perceptive and -interactive model for language, audio, and vision. *arXiv preprint arXiv:2503.01879*, 2025a.

Haotian Liu et al. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Yuanxin Liu et al. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

Zuyan Liu et al. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025b.

Karttikeya Mangalam et al. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023.

Microsoft. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025.

Guangzhi Sun et al. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Changli Tang et al. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Han Wang, Deyi Ji, Lanyun Zhu, Jiebo Luo, and Roy Ka-Wei Lee. Streamsense: Streaming social task detection with selective vision–language model routing. In *Proceedings of the ACM Web Conference 2026*, pp. 8897–8906, 2026a.

Xiaodong Wang, Langling Huang, Zhirong Wu, Xu Zhao, Teng Xu, Xuhong Xia, and Peixi Peng. Livibench: An omnimodal benchmark for interactive livestream video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26517–26525, 2026b. doi: 10.1609/aaai.v40i31.39859.

Yuxuan Wang et al. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. *arXiv preprint arXiv:2503.22952*, 2025.

Boyong Wu et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 28828–28857, 2024. doi: 10.52202/079017-0907.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. *arXiv preprint arXiv:2105.08276*, 2021.

Xiaomi MiMo Team. MiMo-V2.5. `https://mimo.mi.com/docs/en-US/updates/model`, April 2026. Released April 23, 2026.

Jin Xu et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Zhenyu Yang et al. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025.

Yuan Yao et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Hanrong Ye et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Boqiang Zhang et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Peiyuan Zhang et al. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024a.

Yuanhan Zhang et al. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Jiaxing Zhao et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

Jinguo Zhu et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Experiment Implementation Details

Evaluation Settings. The final evaluation set contains 1,421 four-way multiple-choice questions over 470 videos. It comprises 470 near-future prediction questions, 312 perception localization questions, 300 historical understanding questions, 221 interaction dynamics questions, and 118 causal attribution questions. We shuffle the answer choices using a fixed random seed and require each model to return a single option. Empty outputs, unparsable responses, and inference-time errors are scored as incorrect. Omni-modal models receive video with its audio track, vision-only models receive only video frames, and audio-only models receive 16-kHz mono audio. Table 5 summarizes the models and deployment backends used in our evaluation.

API Evaluation Settings. We set `temperature=0` for every API that exposes a temperature parameter to ensure reproducibility. Qwen, Doubao, MiMo, and Gemini are accessed through their official provider APIs. The GPT-series modality-ablation models are accessed through OpenRouter with the provider fixed to OpenAI. Videos are sampled at 1 frame per second (fps).

Local-Model Evaluation Settings. We evaluate local models on a server equipped with seven NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory. Depending on its memory requirements, a model uses one to three GPUs per inference job and remains resident in GPU memory throughout an evaluation run. We use a batch size of 1, the model's native FP16 or BF16 weights without quantization, and FlashAttention-2 whenever supported. All local models configured `do_sample=False`, `temperature=0`. The maximum generation length is 64 tokens, except for Qwen2-Audio, for which we allow 128 tokens because it more frequently produces explanatory text. All models use the same multiple-choice prompt template and the same answer-choice shuffling seed.

For omni-modal models, we adopt an *audio-first, uniform-frame sampling* strategy. We first resample the full audio track to 16-kHz mono and then allocate video frames according to the model's context length and empirically measured GPU-memory capacity. Frames are sampled uniformly over the complete input timeline at no more than 1 fps, rather than truncating the video from the end. When an input exceeds a model's capacity, we temporally compress the audio and video in a synchronized manner or reduce the number of uniformly sampled frames, while preserving coverage of the complete timeline. The maximum numbers of video frames are 128 for Qwen2.5-Omni-7B, 128 for MiniCPM-o-4.5 (77 in the Full setting), 8 for Baichuan-Omni-1.5, 80 for Phi-4-Multimodal, 64 for Ola-7B, 32 for OmniVinci, 286 for HumanOmni-7B, 128 for Nexus-O, and 192 for VITA-1.5. For short videos, the actual number of frames is further bounded by the 1-fps sampling rate.

Phi-4-Multimodal additionally uses token-aware dynamic allocation. We first compute the tokens occupied by the text and complete audio track, estimate the number of image frames that fit in the

Table 5: Models and deployment backends used in the evaluation. † MiMo-V2.5 is evaluated via its official API due to its large parameter size.

Model	Access	Deployment
Qwen3.5-Omni-Flash	Proprietary	Official DashScope API
Qwen3.5-Omni-Plus	Proprietary	Official DashScope API
Doubao-Seed-2.0-Lite	Proprietary	Official Volcengine Ark API
Gemini-2.5-Flash	Proprietary	Official Google Gemini API
Gemini-3.1-Pro-Preview	Proprietary	Official Google Gemini API
MiMo-V2.5†	Open source	Official Xiaomi MiMo API
Qwen2.5-Omni-7B	Open source	Local deployment
MiniCPM-o-4.5	Open source	Local deployment
Baichuan-Omni-1.5	Open source	Local deployment
Phi-4-Multimodal	Open source	Local deployment
Ola-7B	Open source	Local deployment
OmniVinci	Open source	Local deployment
HumanOmni-7B	Open source	Local deployment
Nexus-O	Open source	Local deployment
VITA-1.5	Open source	Local deployment
GPT-4o (vision only)	Proprietary	OpenRouter API
GPT-5 (vision only)	Proprietary	OpenRouter API
GPT Audio	Proprietary	OpenRouter API
GPT Audio Mini	Proprietary	OpenRouter API
Qwen3-VL-8B-Instruct	Open source	Local deployment
Qwen2.5-VL-7B-Instruct	Open source	Local deployment
InternVL3-8B	Open source	Local deployment
VideoLLaMA3-7B	Open source	Local deployment
LLaVA-Video-7B-Qwen2	Open source	Local deployment
MiniCPM-V-2.6	Open source	Local deployment
LongVA-7B	Open source	Local deployment
Qwen2-Audio-7B-Instruct	Open source	Local deployment
Step-Audio-2-mini	Open source	Local deployment
SALMONN-7B	Open source	Local deployment
LTU-AS	Open source	Local deployment

remaining context, and dynamically adjust the memory-based frame limit according to the actual number of visual tokens per frame. HumanOmni uniformly compresses the complete waveform along the temporal axis to fit its 30-second audio window, while VITA uniformly compresses the complete audio features to 4,800 feature steps. Thus, both models retain coverage of the full input duration instead of truncating the audio.

We conduct modality ablations only in the English Local setting. Vision-only models receive no audio. Qwen3-VL-8B, Qwen2.5-VL-7B, and VideoLLaMA3-7B sample at 1 fps with at most 128 frames; InternVL3-8B and MiniCPM-V-2.6 uniformly sample 64 frames from the complete clip; and LLaVA-Video-7B-Qwen2 and LongVA-7B uniformly sample 54 frames. Audio-only models receive no images and operate on 16-kHz mono PCM16 WAV files extracted from the corresponding clips. Qwen2-Audio, Step-Audio-2-mini, and LTU-AS process the complete audio track. Because SALMONN supports an audio window of approximately 30 seconds, we uniformly compress audio longer than 30 seconds to 29.5 seconds rather than truncating its tail. All ablation models likewise use a batch size of 1. If an input exceeds model capacity, we reduce the number of visual frames.

Cost. We report the cost of API inference. Including small-scale feasibility tests and failed retries, the complete evaluation stage costs approximately $450. Agentic question generation, validation, translation, and hardening during dataset construction incur an additional cost of approximately $1,200, resulting in a total API cost of approximately $1,650 for the project.

B Dataset Construction Details

Data Collection and Preprocessing. We collect livestream recordings and platform-provided subtitles from a public video platform. The initial corpus contains 532 videos spanning six scenarios: multi-host PK, chat rooms, outdoor streams, gaming, pets, and commerce. To remove edited compilations, prerecorded broadcasts, and videos without livestream interaction, we uniformly sample five frames from each candidate and use Qwen-VL-Plus to determine whether the visual interface is consistent with a continuous livestream recording and whether observable streamer–audience interaction is present. We manually review borderline cases.

For livestreams longer than 45 minutes, we first use a text LLM (Qwen-Max, selected for its long-context capability) to identify coarse event boundaries and generate summaries from platform-provided subtitles. In our preliminary inspection, such lengthy recordings often contain extended periods of low information density; this step helps identify segments suitable for question construction. We then select a contiguous interval with coherent event progression, substantial audience interaction, and high information density. Shorter livestreams that remain informative throughout are retained in full. All subsequent annotations are performed on the resulting intervals.

After frame-level interaction extraction, we compute for each video the proportion of frames for which no valid audience action is detected (i.e., the extractor returns `[]`). We discard videos whose empty-output rate exceeds 50%. After filtering for video usability, interaction quality, and annotation completeness, 470 videos remain in the benchmark.

Hierarchical Multimodal Representation. For each video, we construct a hierarchical multimodal representation consisting of content semantics and audience interactions. To represent content semantics, we partition each video into non-overlapping 120-second chunks, retaining the final chunk at its actual duration. We use `qwen3-omni-flash-2025-12-01` to jointly process the visual and audio streams of each chunk and generate temporally grounded event segments, visual descriptions, speech and non-speech audio descriptions, and a chunk-level summary.

To represent audience interaction, we sample the video at 1 fps, matching the livestream platform's approximately 1-Hz comment refresh rate. We then use `qwen3.5-flash` to extract usernames, comments, gifts, and other visible interactions, with `temperature=0` and thinking disabled. The extracted actions are aligned at one-second resolution and merged with the platform subtitles and omni-modal descriptions to form a unified timeline.

We use a three-stage highlight-selection procedure to locate question-worthy moments in the long timeline. In the first stage, we construct an audience-activity curve with weights of 3.0, 2.0, 1.0, and 0.6 for comments, gifts, likes, and other actions, respectively. Adaptive short- and long-window peak detection then identifies three to six candidate event segments. In the second and third stages, `qwen3.7-plus` (`temperature=0`, thinking disabled) performs coarse filtering and fine-grained reranking. Coarse filtering receives a compressed timeline of at most 25 seconds for each event segment, whereas reranking receives a detailed timeline of at most 45 seconds for each candidate. The procedure retains one to three highlight candidates per video.

After having obtained these representation, we concatenate the dense description and LLM summarized highlight to generate an overview of the whole livestream as references the subsequent agentic task formation.

Agentic Task Formation. We use an claude code agent powered by GLM-5.2 to generate questions. We select this model based on its quality-cost trade-off in a small-scale pilot study. The annotation workflow is implemented as a fixed skill through which the agent retrieves, on demand, the video overview, candidate highlights, and the subtitles, audience actions, and omni-modal descriptions within a requested time range. The agent therefore does not receive the complete long-video representation in a single context. The specific definition of task definition and livestream category is described in Appendix F

For each sample, the agent first constructs an anchor question whose default type is near-future prediction and then generates diagnostic support questions around its evidence chain. For a near-future prediction question, the agent explicitly defines a prediction cutoff time t. The question context and answer explanation may use only information available before t, and the correct answer must be

supported by at least two independent cues from distinct modalities or information sources. Based on the prerequisite capabilities needed by the anchor, support questions are assigned to perception localization, historical understanding, causal attribution, or interaction dynamics. Every question has four answer choices; distractors are based on interpretations that are present in the context but incomplete or temporally mismatched. Algorithm 1 formalizes this process.

Candidate questions undergo two quality-control tests performed by a three-model test group comprising `google/gemini-3.5-flash`, `qwen/qwen3-235b-a22b-2507`, and `openai/gpt-5-mini`; all test calls use `temperature=0`. In the ***blind test***, a model sees only the question and answer choices, without video context. A question is considered sufficiently discriminative only if at least two models answer incorrectly or fail to provide an answer—equivalently, at most one model may answer correctly. Questions that fail this test are revised by the agent and tested again, for at most three iterations per question.

In the *solvable test*, the models additionally receive the subtitles, audience actions, and omni-modal descriptions within the question's time range, and may answer "not enough information." A question is deemed answerable from the supplied evidence if at least one model answers correctly. If none of the three models answers correctly, whether because they select a wrong option or "not enough information," or because their responses cannot be parsed, the question is discarded. Retained questions additionally pass rule-based checks for structure, temporal bounds, and the causal direction of near-future prediction. Once the Chinese questions are finalized, we use `gemini-2.5-flash` to translate the questions, choices based on the context.

Algorithm. Algorithm 1 details the procedure used to instantiate the agentic task formation pipeline described in Section 3.4. Given the hierarchical representation of a livestream $L = (f, a, i)$, the annotator first examines the global stream context and interaction-burst highlights to locate a prediction point t immediately preceding a salient development. The corresponding prefix $L_{<t}$ then defines the evidence boundary for task construction, such that no information after t is used as input evidence. Based on this prefix, the agent constructs an anchor MCQ q_{anchor} that asks about the evolution of the ongoing situation over the following interval Δ, with all candidate answers constrained to the same decision space $\mathcal{Y}_q$. The anchor is iteratively refined through the blind and solvable tests: the former rejects questions that can be answered without livestream evidence, while the latter verifies that the intended target y^* can be recovered from evidence available in $L_{<t}$. If repeated refinement fails to produce a context-dependent anchor, the agent returns to an alternative prediction point and repeats the procedure.

Once a valid anchor is obtained, the agent traces the evidence and reasoning prerequisites required for its prediction and constructs two to four support questions q_s^j around the same prediction point. Each support question targets one prerequisite capability among perception localization, interaction dynamics, historical understanding, and causal attribution, and is explicitly linked to the anchor through a diagnostic rationale. Supporting evidence may be retrieved from earlier portions of $L_{<t}$ when the relevant causes or interaction history precede the immediate prediction context. Each support question undergoes the same blind and solvable tests and is discarded if it cannot be made both context-dependent and evidence-grounded within the refinement budget. Finally, the validated anchor and support questions are assembled into the complete question set $\mathcal{Q} = \{q_{\text{anchor}}\} \cup \{q_s^j\}$, checked for schema consistency, and translated from the original Chinese tasks into English. This anchor-first construction ensures that the support questions diagnose capabilities directly implicated in the corresponding near-future prediction, rather than forming an independent collection of retrospective understanding tasks.

C Capability Transfer Implementation & Analysis

Dataset. Because collecting and preprocessing additional training data following the same pipeline as in the benchmark was beyond our available budget, we use existing public datasets that provide supervision for capabilities analogous to the four StreamCast diagnostic dimensions. These datasets serve as practical proxies for capability-specific training rather than additional training data collected from StreamCast. We construct the four supervised fine-tuning datasets from public training splits that are independent of StreamCast. For causal attribution (CA) and historical understanding (HU), we use the causal and temporal subsets of NExT-QA, respectively, formulated

Algorithm 1 Anchor-first agentic task formation for one livestream

Require: Livestream $L = (f, a, i)$ with its hierarchical representation H and interaction-burst highlights

Ensure: Validated MCQ set $\mathcal{Q} = \{q_{\text{anchor}}\} \cup \{q_s^j\}$, where each question q has target $y^* = g(q, L_{<t}, \mathcal{Y}_q)$

Phase 1: Locate the prediction point

Read OVERVIEW(L), HIGHLIGHTS(L)
$C \leftarrow H$ ▷ Retrieve the specific context C related to salient development
Choose a primary prediction point t and 1–2 backups from C
Anchor context $\leftarrow L_{<t}$ ▷ all evidence up to t; no signal after t is exposed

Phase 2: Draft and validate the anchor

Draft q_{anchor}: predict how the situation evolves over the following interval Δ, as an MCQ whose options lie in the same decision space $\mathcal{Y}_q$
Self-check: every cited cue lies in $L_{<t}$; no answer leakage; ≥ 2 independent cues
for $r \leftarrow 1$ **to** 5 **do**
    if q_{anchor} passes BLINDTEST($\cdot$) & SOLVABLETEST($\cdot$) **then**
        break ▷ accept: context-dependent and evidence-grounded
    else
        EVOLVE(q_{anchor}) guided by the blind- and solvable-test feedback
    end if
end for
if q_{anchor} is still invalid **then**
    Switch to a backup prediction point t and repeat Phase 2
end if

Phase 3: Build support around the anchor

$P \leftarrow$ prerequisite capabilities of q_{anchor} ▷ perception localization / interaction dynamics / historical understanding / causal attribution
$\mathcal{S} \leftarrow \emptyset$
for each prerequisite $p \in P$ **(2–4 in total) do**
    Select supporting evidence $L_{<t}^{(p)} \subseteq L_{<t}$, tracing back distal causes if needed
    Draft support q_s^j with a `diagnostic_rationale` linking it to q_{anchor}
    for $r \leftarrow 1$ **to** 3 **do**
        if q_s^j passes BLINDTEST($\cdot$) & SOLVABLETEST($\cdot$) **then**
            $\mathcal{S} \leftarrow \mathcal{S} \cup \{q_s^j\}$; **break**
        else
            EVOLVE(q_s^j)
        end if
    end for
    Discard q_s^j if still invalid after 3 rounds
end for
Emit the Chinese tasks, then TRANSLATE($\mathcal{Q}$) into English
return $\mathcal{Q}$

as open-ended video question answering Xiao et al. (2021). For interaction dynamics (ID), we use the context-to-comment-category task from DanmakuTPPBench Jiang et al. (2025). For perception localization (PL), we use the segment-to-event-description task from ActivityNet Captions Krishna et al. (2017), aligning the supervision with describing events in a specified temporal interval. Each single-capability training set contains 1,600 examples to align the data scale for each capability. Each single-capability setting has a separate held-out set of 200 examples for validation, with no video shared between training and held-out evaluation. Since these training datasets involve diverse task formats, including open-ended QA tasks that are difficult to evaluate using accuracy, we uniformly report held-out F1 accuracy in Figure 3 to illustrate training convergence.

To control for training-set size when comparing mixed and single-capability supervision, we construct two mixtures with equal contributions from the four sources. The total-budget-matched setting (`mix4`) contains 400 examples per capability, or 1,600 examples in total, matching the total size of each single-capability training set and controlling for the possibility that a mixture benefits sim-

ply from more training examples. The per-capability-matched setting (`mix4_6400`) contains 1,600 examples per capability, or 6,400 examples in total, matching the amount of supervision for each capability in its corresponding single-capability setting. This second control addresses the possibility that weaker transfer in the smaller mixture results from reducing the number of examples per capability from 1,600 to 400. Together, the two settings control for total training-set size and per-capability training-set size, respectively. Each mixture is evaluated on 200 held-out examples, with 50 from each source. We filter held-out candidates against the union of training-video identifiers across all four sources; this also prevents leakage between the CA and HU subsets, which share NExT-QA as their source dataset. Results in reported in Figure 3 is from setting `mix4_6400`.

Fine-tuning Settings. All settings start independently from Qwen2.5-Omni-7B and use LoRA with rank 8 and scaling factor $\alpha = 16$, while keeping the vision tower frozen. We train for two epochs with a learning rate of 5×10^{-5}, cosine learning-rate decay, a warm-up ratio of 0.1, BF16 precision, a maximum sequence length of 8,192 tokens, and an effective global batch size of eight examples per optimizer update. We report 400 optimizer updates for all four single-capability settings and `mix4`, and 1,600 for `mix4_6400`. Thus, the larger mixture matches the single-capability settings in per-capability data volume and training epochs, but uses four times as many optimizer updates. Experiments use LLaMA-Factory 0.9.3 and Transformers 4.54.1 on an eight-GPU NVIDIA A100 80GB server. Checkpoints are saved every 50 updates and training loss is logged at every update. Results are averaged over three random 3 runs ($p < 0.05$ under t-test).

We measure in-domain learning using generated-answer token-F1 on the held-out source data, separately from multiple-choice accuracy on StreamCast. Audio use in mixed-source held-out evaluation follows the corresponding single-source setting; audio is disabled for DanmakuTPP samples, whose videos lack audio tracks. For transfer evaluation, we use the English clip setting and the same StreamCast evaluation pipeline for the base model and all adapters. The evaluation contains 470 future prediction (FP), 118 CA, 300 HU, 221 ID, and 312 PL questions. Their base-model accuracies are 32.77%, 78.81%, 69.00%, 65.61%, and 70.51%, respectively. The standard-size settings are evaluated at 50-update intervals through update 400; the larger mixture is evaluated every 200 updates through update 1,600. The base model provides the zero-update reference. Only CA, HU, and ID at the PL adapter's update-50 checkpoint are missing from the cross-dimension evaluations. Its FP and PL evaluations are complete.

Cross-Dimension Transfer. The transfer patterns in Figure 3 distinguish learning on the auxiliary source tasks from transfer to StreamCast. CA and HU supervision yields positive peak gains on their corresponding diagnostic dimensions and on FP, with broader improvements across the remaining dimensions. In contrast, ID and PL achieve substantial held-out learning progress on their source tasks, with token-F1 increasing from 8.5% to 52.2% and from 33.1% to 62.6%, respectively, but show only limited improvement on the corresponding StreamCast diagnostic dimensions. Their FP gains are similarly small. This suggests that the limited prediction transfer primarily coincides with weak transfer from the auxiliary supervision to the target diagnostic capability, rather than with successful diagnostic improvement failing to benefit prediction.

One likely factor is the imperfect correspondence between the auxiliary tasks and the diagnostic dimensions in StreamCast. Although we deliberately selected tasks that most closely reflect each target capability, the public datasets provide only approximate supervision. ID is trained from a DanmakuTPPBench in which danmaku functions largely as post-hoc commentary on video content, rather than as real-time audience interaction that can influence the evolution of a livestream. For PL, ActivityNet Captions provides segment-level event descriptions that emphasize temporal localization, which only partially overlaps with the fine-grained localization of visual evidence required in StreamCast. Consequently, strong learning on the source task does not guarantee comparable improvement on the corresponding StreamCast support questions.

Mixed supervision provides a complementary comparison. The total-budget-matched and per-capability-matched mixtures yield peak FP gains of +1.70% and +2.55% percentage point, respectively. Increasing supervision improves observed transfer, but neither mixture exceeds CA or HU training alone. Together, these results suggest that prediction gains broadly track whether auxiliary training transfers to the diagnostic dimensions themselves, while the proxy nature of the training data prevents attributing differences solely to the intrinsic importance of individual capabilities.

Table 6: Anchor accuracy grouped by performance on the corresponding support questions in the English Local setting.

Matched support accuracy	Anchor accuracy
0%	31.1%
$(0, 50]\%$	37.6%
$(50, 100)\%$	44.6%
100%	45.2%

All heatmap entries report peak accuracy changes relative to the same base model, computed independently for each evaluation dimension over the evaluated checkpoints; gains within the same row therefore need not be attained simultaneously by a single checkpoint. Results are averaged over three random runs, with statistically significant improvements reported under the t-test ($p < 0.05$) setting described above, indicating that the observed transfer patterns are not driven by a single training run. Nevertheless, the auxiliary public datasets serve only as proxies for the StreamCast diagnostic dimensions and differ from the benchmark in task formulation, domain, and video distribution. Therefore, differences in transfer strength should not be interpreted as establishing a causal hierarchy among the four capabilities. In particular, weak transfer for ID and PL may reflect limited transfer from the auxiliary supervision to the corresponding StreamCast diagnostic dimensions, rather than evidence that these capabilities are intrinsically less relevant to future prediction.

D Diagnostic Value Analysis

In StreamCast, support questions are constructed from the evidence chain of each anchor, however, this construction alone process does not guarantee their diagnostic value. In addition to the human evaluation in 5.1, we further conduct a statistical test of whether a model's performance on the support questions is informative of its performance on the corresponding anchor, beyond its general capability across the benchmark.

We analyze the anchor observations from 15 models in the English Local setting, where each question is evaluated independently. For each anchor, we compute the proportion of its matched support questions answered correctly. As shown in Table 6, anchor accuracy increases from 31.1% when all matched supports are answered incorrectly to 45.2% when all are answered correctly. We then fit a logistic regression controlling for model identity, livestream category, and video duration. To determine whether the association is specific to the constructed pairing, we randomly reassign support results among comparable questions within the same model, category, and support type, and repeat the analysis 1,000 times.

The coefficient for the original pairing is $\beta = 0.204$, compared with a mean of 0.028 and a 95% interval of $[-0.121, 0.183]$ under random pairing (one-sided permutation $p = 0.013$), indicating that the matched support questions contain anchor-specific diagnostic information that is not explained solely by general differences across models or support types. The result supports their use for diagnosing prediction failures, but does not establish that the four capabilities causally determine future prediction.

E Human Evaluation of Language Fidelity

Because StreamCast is constructed from Chinese livestreams while the main experiments are conducted using English questions, we further examine whether translation alters the semantic content or difficulty measured by the benchmark. Using the same set of sampled livestreams and annotators as in the quality analysis, we evaluate the Chinese-to-English translations along four criteria: *semantic equivalence*, *difficulty preservation*, *cultural preservation*, and *pragmatic fidelity*. Detailed rating guidelines are provided in Table 8.

As shown in Table 7, semantic equivalence and difficulty preservation receive near-ceiling ratings of 2.99 and 2.98, respectively, indicating that the translations largely preserve both the intended meaning and the reasoning requirements of the original questions. Cultural preservation and pragmatic fidelity remain high at 2.90, with the remaining differences primarily arising from livestream-

Table 7: Human assessment of StreamCast questions for translation fidelity. For each criterion we report the mean score and the inter-annotator consistency (pairwise exact-agreement rate).

Criterion	*Semantic Equivalence*	*Difficulty Preservation*	*Cultural Preservation*	*Pragmatic Fidelity*
Mean	2.99	2.98	2.90	2.90
Consistency (%)	98.3	95.0	81.7	80.0

specific expressions, platform terminology, and colloquial language. These results suggest that the English questions faithfully preserve the core content and task requirements of the original Chinese questions, while some culturally grounded nuances may be attenuated during translation.

This finding complements the language alignment analysis in the main paper. The observed performance differences between Chinese and English prompts are therefore unlikely to arise primarily from substantial changes in question semantics, but may instead reflect differences in how models process and reason over language-aligned livestream content.

F Task Definitions

F.1 Livestream Categories

StreamCast covers six livestream categories based on common content types on the source platform. The categories are distinguished by their dominant content, interaction patterns, and the types of multimodal understanding they involve. Specific examples of different types of livestreams are illustrated in Figure 4.

Multi-host PK. Multi-host PK livestreams involve two or more streamers participating in competitive activities, often governed by platform-specific mechanisms such as audience gifts and rankings. Understanding these streams requires tracking interactions among streamers, interpreting audience participation, and relating platform feedback to the progression of the competition.

Chat Rooms. Chat-room livestreams primarily feature one or more streamers engaged in conversation with each other or their audiences. Their content is largely conveyed through speech, requiring models to follow narrated events, maintain conversational context, and interpret responses to audience comments.

Outdoor. Outdoor livestreams take place in physical environments beyond a fixed, streamer-centered view and may involve movement, changing scenes, and interactions with surrounding people or objects. Understanding these streams requires integrating visual observations with speech and other contextual signals to track developments across locations and events.

Gaming. Gaming livestreams feature streamers playing video games, with visual content primarily consisting of virtual environments and entities. Understanding these streams requires interpreting game states, actions, and rules while relating in-game developments to the streamer's commentary and audience interactions.

Pets. Pet livestreams primarily feature animals and their activities. Understanding these streams requires tracking animal behavior, recognizing interactions among animals or between animals and humans, and interpreting developments that may not be explicitly described through speech.

E-commerce. E-commerce livestreams focus on product presentation and sales. Understanding these streams requires relating product demonstrations, promotional information, and sales strategies to audience questions, streamer responses, and subsequent developments.

F.2 Question Types

StreamCast contains five question types: one anchor type targeting near-future prediction and four support types assessing related diagnostic capabilities. The anchor question targets an outcome in

Table 8: Human evaluation criteria and scoring guidelines.

Criterion	Score 3	Score 2	Score 1
	Quality Assessment		
Relevance	The question and options are relevant to the video content, and the answer corresponds to the correct temporal segment.	The question is generally relevant, with minor deviations in content or temporal grounding that do not affect judgment.	The question contradicts or fabricates video content, or the answer corresponds to an incorrect temporal segment.
Evidence Support	The reference explanation sufficiently and directly supports the reference answer.	The explanation is relevant but provides relatively weak support for the reference answer.	The explanation cannot support the reference answer or conflicts with it.
Unambiguity	There is only one correct answer, and distractors are reasonable and unambiguous.	There is only one correct answer, but some distractors are weak or potentially confusing.	Multiple answers could be correct, or the correct answer is not clearly identifiable.
Difficulty	The difficulty is appropriate and requires multimodal reasoning.	The question is slightly too easy or difficult but remains answerable.	The question is overly simple or excessively difficult, e.g., unanswerable or beyond the available evidence.
Modality Variety	Answering the questions requires information from at least two modalities, such as visual content, audio, or audience interaction.	The questions can be answered using only a single modality.	The questions can be answered without relying on video information, e.g., using only the question itself or common knowledge.
Supportiveness	The questions are strongly associated with the anchor question and effectively support the diagnosis of relevant abilities.	The questions are related to the anchor question but provide limited diagnostic support.	The questions are disconnected from the anchor question and cannot support ability diagnosis.
	Language Assessment		
Semantic Equivalence	The question and all options preserve the original semantics and distractiveness.	The semantics are generally preserved, with minor changes in distractiveness that do not affect answering.	The semantics of the question or key options are altered, causing the correct answer to change.
Difficulty Preservation	The correct answer remains unchanged and the difficulty is preserved.	The translation causes minor changes in difficulty.	The translated question becomes unanswerable or directly reveals the answer.
Cultural Preservation	Cultural references, livestream-specific expressions, platform contexts, and proper nouns are appropriately preserved.	Most cultural information is preserved, but some entities are awkwardly translated or omitted without affecting comprehension.	Important cultural contexts or entities are substantially lost or mistranslated.
Pragmatic Fidelity	The tone, intent, and colloquial style of the original expression are naturally preserved.	The original pragmatic meaning is generally preserved, but the style becomes slightly formal or flat.	The tone, sarcasm, interaction intent, or other pragmatic meaning is lost, or the expression is clearly unnatural.

the immediate future, whereas support questions examine information and reasoning relevant to understanding the preceding stream. Specific examples of different types of questions are listed in Figure 4.

Future Prediction. Future prediction questions ask models to infer how an ongoing situation will develop immediately after a designated prediction point. The prediction must be grounded in preceding multimodal evidence, including relevant events, interaction patterns, and contextual cues. Candidate answers represent plausible outcomes within the same decision space, so the task requires distinguishing among possible developments rather than relying on superficial differences between options.

Perception Localization. Perception localization questions assess whether models can identify and locate specific visual, or acoustic evidence within the preceding livestream. They target the recognition of relevant entities, states, or signals in the video content and their association with the appropriate temporal context.

Interaction Dynamics. Interaction dynamics questions assess whether models can understand how streamers, audiences, and other participants respond to one another. Questions of this type focus on relationships between actions and reactions, including responses conveyed through speech, visible behavior, comments, gifts, or other interaction signals.

Historical Understanding. Historical understanding questions assess whether models can track events and maintain relevant context across earlier portions of a livestream. It involves identifying event sequences, temporal relationships, and developments that connect earlier occurrences to the current situation.

Causal Attribution. Causal attribution questions assess whether models can identify evidence-supported explanations for observed events, decisions, or reactions. They require connecting a development to relevant preceding actions, statements, or contextual changes.

G LIMITATIONS

There are several directions in which StreamCast can be further improved:

- **First, the benchmark construction pipeline still relies substantially on model-assisted annotation.** Although we incorporate rule-based constraints, multi-model blind and solvable tests, and human quality assessment to improve the reliability of the generated questions, these mechanisms cannot completely eliminate annotation errors or systematic model biases. In particular, multiple stages of the pipeline, including multimodal description, speech recognition, audience-action extraction, highlight selection, question generation, verification, and translation, involve model-generated intermediate results. Errors introduced at an earlier stage may therefore propagate to subsequent stages. Moreover, the generation and automatic verification procedures primarily operate on textual representations derived from the livestream rather than directly examining the raw video, potentially allowing questions to remain internally consistent with the intermediate annotations while being imperfectly grounded in the original content. Although we have conducted human assessment to further validate the overall quality of StreamCast, due to budget limitation, it covers only a sampled subset of the benchmark and does not constitute exhaustive verification. Future work could incorporate broader human-in-the-loop review and direct video-grounded verification to further improve annotation reliability.
- **Secondly, the current benchmark has limited platform and content coverage.** All livestreams in StreamCast are collected from Bilibili and are therefore influenced by the platform's language, interface design, cultural conventions, and audience-interaction mechanisms. The distribution across livestream categories is also imbalanced, with multi-host PK and chat-room streams accounting for a substantial proportion of the corpus. In addition, our data filtering and highlight-selection procedures favor interaction-rich and event-dense intervals, since audience-activity bursts are used as an initial signal for locating question-worthy moments. Consequently, quieter livestreams, in which relevant developments unfold gradually or receive limited audience reaction, may be underrepresented. These characteristics may limit the direct generalizability of the benchmark to other platforms, languages, cultural contexts, and livestream genres. In future work, we plan to expand StreamCast with data from more diverse platforms and regions, while improving the balance across content categories and interaction patterns.



Category: Multi-host PK **Anchor:** Future Prediction

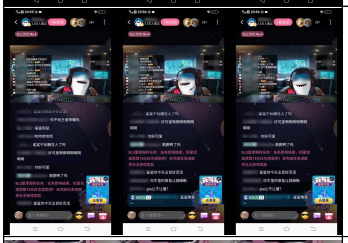
Question: A streamer is continuously and actively initiating random PK connections (随机PK连麦) during a livestream. Immediately after, when a new connection is established and the other party asks "how do you want to play?" (怎么玩), what will the streamer most likely do first?

A. Apologize and say they just wanted to chat casually and do not really want to play, without actually engaging in a battle. ✅

B. Respond enthusiastically to the other party and actively initiate a PK talent battle.

C. Invite the other party to sing a duet/co-sing a song together.

D. Ask the other party what talents they are good at and agree on a PK wager.



Category: Game **Anchor**: Future Prediction

Question: Starting from 00:00:00, the streamer is playing games like Fall Guys. After a round of Fall Guys ends, the streamer takes an interest in another game, only to find that it only has a mobile version and cannot be played on a computer. The streamer says "can't download it" (下不了) and says "goodnight" (晚安) to the audience. Based on the streamer's situation, words, and actions at this moment, what is he most likely to do next?

A. Repeatedly attempt to download that mobile-only game on the computer.

B. End the livestream right here and finish tonight's broadcast.

C. Switch to using a mobile phone to download and try out that mobile-only game.

D. Stay live, continue the broadcast, and transition to chatting with the audience about the next game to play. ✅



Category: Pet **Support**: Perception Localization

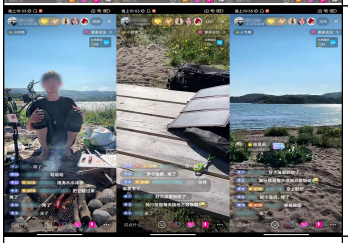
Question: At the end of the video, when the off-screen narration explains that the cat is confused because of its "rabbit costume," which option best matches the shot language and the visual details of the cat?

A. The camera slowly pushes in to a close-up; the cat raises its head and looks directly at the camera, with both ears symmetrically flattened backward.

B. The camera remains in a fixed high-angle medium shot; the cat looks diagonally downward, and the two ears of the rabbit hat tilt asymmetrically, one forward and one back. ✅

C. The camera pans right; the hat is severely tilted and covers one of the cat's eyes.

D. The camera sways left and right at eye level; the cat lowers its head to sniff the ground, with one hat ear folded forward and the other fully upright.



Category: Outdoor **Support**: Causal Attribution

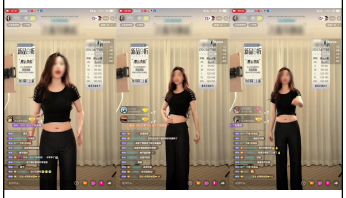
Question: In this outdoor beachcombing (赶海) livestream, after processing a batch of crabs by the campfire, the streamer asked the audience "should we keep going?" and mumbled about adding more fire to continue boiling them. Regarding his actual handling of "continuing to boil the crabs" next, which of the following is correct?

A. He packed the crabs into his bag to take away, intending to save them to boil another day.

B. He made the fire burn brighter and continued to boil and eat the remaining bag of crabs.

C. He felt the fire was not strong enough, so he simply packed up the campfire and stopped boiling.

D. He originally wanted to add fire to continue boiling, but judged that he was already full, so he gave up, did not add more fire, and did not continue boiling. ✅



Category: E-commerce **Support**: Interaction Dynamics

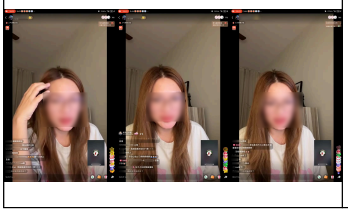
Question: In the second half of this livestream video, while the host was introducing these pants, the bullet comments (弹幕) scrolled densely. Which word was the most frequently sent positive review by the audience?

A. Slimming (显瘦).

B. Pilling (起球).

C. Comfortable to wear (好穿). ✅

D. Amazing (绝了).



Category: Chatroom **Support**: Historical Understanding

Question: The streamer mentioned in the second half of the livestream that the leader evaluated this male colleague as "a pitfall" (是个坑). According to her narrative, when and on what occasion did the leader say this?

A. At the male colleague's wedding, the leader said it in public.

B. After the male colleague had already quickly registered for marriage, the streamer heard it from his leader during a dinner gathering with friends. ✅

C. When the streamer proactively went to ask the leader "how is this person?", the leader told her.

D. When the male colleague was still asking around for matchmaking introductions and had not yet registered for marriage, the streamer heard the company leader evaluate him in the office.

Figure 4: Examples of StreamCast tasks

- **In addition, our multiple-choice formulation simplifies the inherently uncertain nature of future prediction.** StreamCast constructs each prediction task around a retrospectively selected prediction point and treats the outcome that actually occurred as the unique correct answer. This design provides an objective and reproducible evaluation protocol, but a given livestream prefix may support multiple plausible continuations in real-world scenarios. The benchmark does not explicitly measure whether a model can represent such uncertainty, assign calibrated probabilities to alternative outcomes, or generate possible futures beyond the predefined candidate set. Moreover, because prediction points are selected after observing the complete livestream and are

concentrated around salient developments, the resulting tasks may exhibit hindsight and event-selection biases. Therefore, StreamCast should primarily be interpreted as evaluating evidence-grounded near-future outcome selection rather than fully open-ended or probabilistic forecasting. Future work could introduce naturally sampled prediction points, multiple acceptable outcomes, and probability-based evaluation protocols to better reflect the uncertainty of real-world livestream evolution.

- **Lastly, the current evaluation remains different from fully online livestream understanding.** Although StreamCast is constructed from livestream recordings and strictly excludes information after the prediction point, the evaluated models receive a preassembled local clip or complete video prefix in an offline setting. The current protocol therefore does not assess several requirements of practical streaming systems, such as incrementally updating the model state as new observations arrive, proactively determining when a prediction should be made, maintaining long-term memory over continuously evolving streams, or satisfying explicit latency and computation constraints. Furthermore, evaluated models differ substantially in their supported context lengths and input-processing mechanisms. Variations in frame sampling, audio compression, context capacity, and API stability may affect Full-context performance in addition to the models' underlying perception and reasoning abilities. Future work could extend StreamCast toward a standardized online protocol in which models process multimodal inputs incrementally and make predictions under controlled memory, latency, and computational budgets.