# Live-Aid: A Large-Scale Dialogue Dataset and Benchmark for Interleaved Multi-party Interactions in Live Streaming


**Yiming Lei[1§], Yize Fan[1§], Zeming Liu[1†], Jiaji Dong[2], Hui Qiu[2♠], Haitao Leng[2], Qingjie Liu[1†], Kehai Chen[3], Tingting Gao[2], Yunhong Wang[1],**

[1]School of Computer Science and Engineering, Beihang University
[2]Kuaishou Technology, [3]Harbin Institute of Technology, Shenzhen

[†]Corresponding authors [♠]Project Leader

ymlei@buaa.edu.cn, zmliu@buaa.edu.cn



## Abstract

Recent advancements in Multimodal Large Language Models (MLLMs) have achieved significant success in understanding static pre-recorded video scenarios (e.g., event-centric or narrative-driven content). However, existing MLLMs are largely trained on datasets restricted to static content due to the scarcity of high-quality interleaved data, causing them to struggle with dynamic interactions. Distinct from pre-recorded videos, live streaming is characterized by high-density, interleaved multimodal turns, where viewer comments (*danmaku*) are tightly coupled with real-time audio-visual evidence and evolving dialogue context. In such settings, purely textual annotations fail to capture fine-grained visual and temporal dependencies. To bridge this gap, we introduce **Live-Aid**, the first large-scale interleaved live interaction Chinese dataset with **human-annotated**, temporally aligned video responses, spanning over **1,100 hours** and 80,037 dialogue turns across 8,053 video sessions. Building on this, we leverage these high-quality annotations within a novel multi-agent pipeline to construct evaluation tasks targeting core capabilities of live interactions. Extensive evaluations of strong Video-LLMs and Omni-LLMs reveal critical limitations in interleaved multi-turn interactions requiring temporal reasoning, highlighting the value of **Live-Aid** in advancing interleaved multimodal reasoning and dynamic audio-visual dependencies. Live-Aid is publicly available at this project repository.



Figure 1: **An interleaved dialogue example from Live-Aid.** Host-Danmaku interactions are annotated with temporal dependencies: *Instant*, *Delay*, and *History*.

## 1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) have made remarkable strides in the field of multimodal dialogue and understanding (Hurst et al., 2024; Shu et al., 2025; Bai et al., 2025b; Tang et al., 2025). Building upon these advancements, a range of benchmarks (Li et al., 2024a; Fu et al., 2025; Qiang et al., 2025) have been established to evaluate the capabilities of these models, primarily through question-answering (QA) tasks. While previous works (Yi et al., 2024) focus on static, pre-recorded videos that are event or narrative driven, these datasets are largely constrained to passive observation.

Distinct from pre-recorded videos and such passive observation, live streaming (e.g., e-commerce) epitomizes a dynamic environment characterized by high information density and continuous ***viewer-host interaction*** via **interleaved multimodal turns**. For example, timestamped live **danmaku** (*i.e., a form of real-time on-screen viewer commentary overlaid on the video*) are tightly synchronized with real-time *audio-visual evidence*, creating intricate interactions rich in multimodal dynamics, as illustrated in Fig. 1. In this context, simple textual replies are often insufficient, whereas direct video responses offer significantly clearer visual grounding and vividness. However, most current datasets

§ indicates that this work was done during an internship at Kuaishou Technology.




*Findings of the Association for Computational Linguistics: ACL 2026*, pages 23813–23850

limit these interactions to static textual annotations, failing to capture the high-quality interleaved data and the audio-visual evidence essential for such direct responses.

To address these challenges, we introduce **Live-Aid**, the first large-scale, human-annotated collection of danmaku and video responses derived from Chinese live streams for interleaved live stream interaction. To ensure superior data quality, we employed 10 expert annotators and establish a rigorous annotation pipeline. Distinct from previous coarse-grained labeling, our protocol mandates fine-grained temporal grounding within the dialogue context: annotators are required to explicitly align the *danmaku* queries and their corresponding host response clips with precise audio-visual spans, thereby bridging the gap between abstract conversational reasoning and concrete evidence. In total, **Live-Aid** spans over **1,100 hours** across 1,763 live streams, comprising 8,053 video sessions covering 44 diverse e-commerce categories characterized by high-density interactions.

To comprehensively evaluate interleaved live interaction, we focus on three fundamental capabilities essential for dynamic stream environments (Chen et al., 2024a): (1) **[Multimodal Understanding]**, which involves deciphering cross-modal alignment and viewer intents; (2) **[Dialogue Modeling]**, which entails simulating social interactions and summarizing context; and (3) **[Temporal Reasoning]**, a critical skill that requires models to navigate the streaming timeline to resolve cross-turn dependencies and ground responses to specific timestamps. Drawing inspiration from recent advancements in agent-enhanced dataset synthesis (Sun et al., 2025; Jiang et al., 2025), we leverage our high-quality annotations within a multi-agent collaborative framework to operationalize these capabilities into eight core evaluation tasks. Specifically, we orchestrate four specialized agent groups to enrich raw annotations with dense visual captions and construct challenging adversarial components. To guarantee the rigorousness of **Live-Aid**, we employ a Consensus Committee mechanism combined with human expert review, strictly filtering tasks for solvability and correctness. Extensive evaluations on 17 leading Video-LLMs and Omni-LLMs reveal significant limitations in processing high-density interleaved multimodal inputs and complex temporal dependencies, highlighting the necessity of interleaved multimodal perception and temporal reasoning for authentic live interaction. We summarize our contributions as follows:

**(1) Novel Task for Interleaved Live Interaction.** We define the task of interleaved live stream interaction, shifting from passive understanding of static pre-recorded content to active interactions in dynamic live streaming scenarios.

**(2) Comprehensive Live Interaction Resource.** We introduce *Live-Aid*, the first large-scale resource comprising 8,053 video sessions (1,100+ hours) of human-annotated interleaved live dialogues, accompanied by an agent-enhanced benchmark for rigorously evaluating dynamic live scenarios.

**(3) Extensive Experiments and Evaluation.** We conduct extensive evaluations on *Live-Aid*, revealing significant limitations of existing models and guiding future research in handling high-density interleaved dynamics and temporal dependencies.

## 2 Related Work

**Multimodal Dialogue.** Prior research on text-based dialogue has mainly focused on applications such as conversational recommendation and user goal clarification (Liu et al., 2020, 2021, 2022). Early multimodal dialogue evolved from single-turn image-text tasks like VQA v2.0 (Goyal et al., 2017) to multi-turn conversations represented by VisDial (Das et al., 2017) and MMChat (Zheng et al., 2022). To further enhance multimodal interactivity, subsequent studies introduced interleaved image-text paradigms, as seen in PhotoChat (Zang et al., 2021) and DialogCC (Lee et al., 2024). Recent research has transitioned towards the video domain, shifting focus from event-centric QA (e.g., ActivityNet-QA (Yu et al., 2019) and VDAct (Imrattanatrai et al., 2025)) to multi-party social dialogues that incorporate user comments (Lin et al., 2023; Shi et al., 2025). While LiveChat (Gao et al., 2023) and DanmakuTPP (Jiang et al., 2025) extend this line of research to large-scale danmaku, they largely treat comments as static metadata and overlook their real-time interaction with the video stream. As a result, existing "watch-then-talk" paradigms still struggle to capture the interleaved dynamics of live streaming and the fine-grained temporal alignment between streams and danmaku.

**Video-LLMs and Evaluation.** The rapid advancement of Large Language Models (LLMs) (Achiam et al., 2023; Guo et al., 2025; Comanici et al., 2025; Yang et al., 2025a) has accelerated the research transition from text to visual perception. Video-LLMs have progressed from early temporal mod-

| Datasets | Response | V.I. | M.P. | Temp. | Source | Anno. | Capability | | | #Visual | #Dmk | #Turns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | MU | DM | TR | | | |
| VisDial (Das et al., 2017) | Text | ✗ | ✗ | ✗ | Image | Human | (✓) | (✓) | ✗ | 123K | - | 2.47M |
| Image-Chat (Shuster et al., 2020) | Text | ✗ | ✗ | ✗ | Image | Human | (✓) | (✓) | ✗ | 202K | - | 401K |
| OpenViDial2.0 (Wang et al., 2021) | Text | ✗ | ✗ | ✗ | Image | Rule-filtered | (✓) | (✓) | ✗ | 5.6M | - | 5.6M |
| PhotoChat (Zang et al., 2021) | Image, Text | (✓) | ✗ | ✗ | Image | Human | (✓) | (✓) | ✗ | 10.9K | - | 117K |
| MMChat (Zheng et al., 2022) | Text | ✗ | ✗ | ✗ | Image | Rule-filtered | (✓) | (✓) | ✗ | 204K | - | 314K |
| MMDialog (Feng et al., 2023) | Image, Text | (✓) | ✗ | ✗ | Image | Rule-filtered | (✓) | (✓) | ✗ | 1.53M | - | 4.92M |
| DialogCC (Lee et al., 2024) | Image, Text | (✓) | ✗ | ✗ | Image | LLM | (✓) | (✓) | ✗ | 129K | - | 929K |
| MUSE (Wang et al., 2025c) | Image, Text | (✓) | ✗ | ✗ | Image | Multi-agent | (✓) | (✓) | ✗ | 13.7K | - | 83K |
| TikTalk (Lin et al., 2023) | Text | ✗ | ✓ | ✗ | Video | Rule-filtered | (✓) | (✓) | ✗ | 38.7K | ✗ | 827K |
| LiveChat (Gao et al., 2023) | Text | ✗ | ✓ | ✗ | Live Stream | Rule-filtered | (✓) | (✓) | ✗ | 182K | 1.3M | 1.3M |
| MMLSCU (Meng et al., 2024) | Text | ✗ | ✓ | ✓ | Live Stream | Human | (✓) | ✗ | ✗ | 0.2K | 50.1K | ✗ |
| KwaiChat (Shi et al., 2025) | Text | ✗ | ✓ | ✗ | Video | Rule-filtered | ✗ | (✓) | ✗ | 93.2K | ✗ | 0.8M |
| Friends-MMC (Wang et al., 2025b) | Image, Text | (✓) | ✓ | ✓ | Video | Rule-filtered | (✓) | (✓) | ✗ | 15.6K | ✗ | 24k |
| DanmakuTPP-Events (Jiang et al., 2025) | ✗ | ✗ | ✓ | ✓ | Video | ✗ | ✗ | (✓) | ✓ | 7.2K | 10.8M | ✗ |
| TikTalkCoref (Li et al., 2025b) | Image, Text | ✗ | ✗ | ✓ | Video | Human | (✓) | ✗ | ✗ | 1.1K | ✗ | 2.2K |
| MVBench (Li et al., 2024a) | Text | ✗ | ✗ | ✗ | Video | LLM | (✓) | ✗ | ✗ | 3.6K | ✗ | 8.0K |
| MMBench-Video (Fang et al., 2024) | Text | ✗ | ✗ | ✗ | Video | Human | (✓) | ✗ | ✗ | 0.61K | ✗ | 4.0K |
| AutoEval-Video (Chen et al., 2024b) | Text | ✗ | ✗ | ✗ | Video | Human | (✓) | ✗ | ✗ | 0.33K | ✗ | 15.1K |
| VideoVista (Li et al., 2024b) | Text | ✗ | ✗ | ✗ | Video | LLM | (✓) | ✗ | (✓) | 3.4K | ✗ | 49.8K |
| Video-MME (Fu et al., 2025) | Text | ✗ | ✗ | ✗ | Video | Human | (✓) | ✗ | ✗ | 0.90K | ✗ | 5.4K |
| OVBench (Huang et al., 2025) | Text | ✗ | ✗ | ✓ | Video | LLM | (✓) | ✗ | ✓ | 1.5K | ✗ | 14.2K |
| OVOBench (Niu et al., 2025) | Text | ✗ | ✗ | ✓ | Video | LLM, Human | (✓) | ✗ | ✓ | 0.64K | ✗ | 5.6K |
| **Live-Aid (Ours)** | Video, Audio, Text | ✓ | ✓ | ✓ | Live Stream | Human | ✓ | ✓ | ✓ | **8.1K** | **53.3K** | **80.0K** |

Table 1: **Comparison between Live-Aid and existing multimodal datasets.** ***Modalities***: Text, Image, Video, Audio, Live Stream. ***Anno.*** represents the annotation method: Human, LLM, Rule-filtered, Multi-agent collaboration. ***V.I.*** denotes Video–Text Interleaved sequences; ***M.P.*** indicates multi-party interactions; ***Temp.*** represents temporal alignment. ***MU***, ***DM***, and ***TR*** denote Multimodal Understanding, Dialogue Modeling, and Temporal Reasoning capabilities of live interaction in Sec. 4.1, respectively. ***#Visual***, ***#Dmk***, and ***#Turns*** are counts of visuals, danmaku, and turns.

eling (Zhang et al., 2023; Li et al., 2025a) to advanced unified architectures like Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025a), while emerging Omni-models (Xu et al., 2025; Ye et al., 2025; OpenAI, 2025a) further enable simultaneous audio-visual perception. To assess these capabilities, research community has established diverse benchmarks (Li et al., 2024a; Fu et al., 2025) for video comprehension, and has gradually extended to streaming video understanding (Huang et al., 2025; Niu et al., 2025). However, these benchmarks mostly rely on generic video sources, overlooking the distinctive interleaved interactions between live streams and viewer comments. To bridge this gap, we construct Live-Aid, a large-scale dataset for interleaved live dialogue interaction, designed to support the study of core capabilities in dynamic live-streaming scenarios.

## 3 Live-Aid Dialogue Construction

To address the scarcity of high-quality resources capturing vivid audio-visual evidence within high-density interleaved dynamics, we present **Live-Aid**, the first large-scale interleaved dialogue dataset with fine-grained temporal annotations. In this section, we detail our design specifications and the rigorous construction pipeline.

### 3.1 Problem Formulation

A live session is denoted as $\mathcal{S} = (\mathcal{V}, \mathcal{A}, \mathcal{D})$, where $\mathcal{V}, \mathcal{A}$ are continuous visual and audio streams, and $\mathcal{D}$ is the ordered danmaku sequence. We model the interaction as a sequence of turns $\mathcal{S}_{dial} = \{T_k\}_{k=1}^{M}$. Each turn is defined as a triplet $T_k = (\mathcal{C}_k, d_k^*, \mathbf{v}_k)$, where: 1) $\mathcal{C}_k \subset \mathcal{D}$ is the ***contextual pool*** of ambient comments within a temporal window; 2) $d_k^* \in \mathcal{C}_k$ is the ***target danmaku*** triggering the response; and 3) $\mathbf{v}_k$ is the ***multimodal response*** segment $\mathcal{V}[t_s : t_e] \oplus \mathcal{A}[t_s : t_e]$. The intrinsic structure of live interactions is defined as the evolution of each turn $T_k$ conditioned on history $H_{k-1}$, reflecting the tight coupling between sparse user queries and continuous streaming evidence.

### 3.2 Dataset Construction

**Data Collection.** We collect data from *Kuaishou*[1], a leading platform with a vast user base. Focusing on e-commerce streams, we capture dense interactions, rich multimodal contexts, and broad domain

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

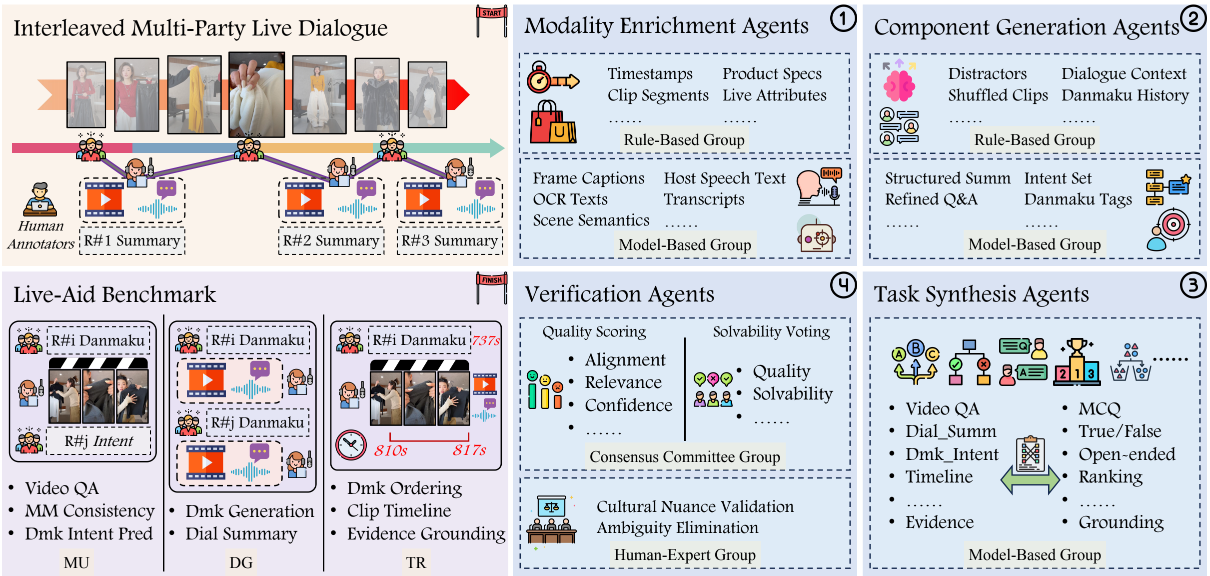



Figure 2: **Multi-Agent Collaborative Generation Framework for constructing Live-Aid.** The framework comprises four agent groups: (1) Modality Enrichment extracts basic dialogue metadata and performs temporal alignment and visual captioning; (2) Component Generation constructs fundamental blocks such as adversarial distractors and dialogue contexts; (3) Task Synthesis utilizes these components to synthesize diverse tasks across Multimodal Understanding (MU), Dialogue Modeling (DM), and Temporal Reasoning (TR); and (4) Verification Agents employ a Consensus Committee and human expert review to guarantee solvability and quality.

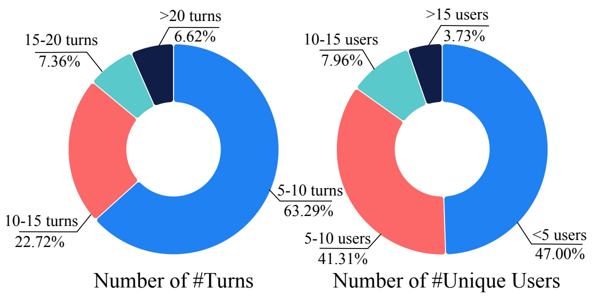



Figure 3: **Distribution of turns and users in Live-Aid.**

coverage. The raw dataset covers 59 categories (e.g., Clothing, Beauty, Health), comprising 12.7K live streams and 343.6K danmaku comments.

**Data Cleaning.** Raw live streams often contain significant redundancy (e.g., silence, waiting periods). As a result, we designed a rigorous three-stage cleaning pipeline to ensure high data quality. First, we used ASR and visual change detection to segment long streams into coherent, product-centric clips. Next, we applied a multi-level danmaku filtering strategy, combining rule-based heuristics and intent detection models to extract genuine information-seeking queries while removing spam and toxic content. Finally, to support effective multimodal modeling, we curated the dataset under strict quality constraints—retaining only sessions with clear audio, appropriate duration (1-20 min), and realistic multi-party interaction density (4-30 unique users). This process produced 20,406 high-quality sessions from the raw collection.

**Data Annotation and Quality Control.** To address the challenge of fine-grained temporal alignment between textual queries and dynamic audio-visual responses, we employed a team of ***10 professional annotators*** following a rigorous protocol. During this phase, annotators served as a final quality filter, explicitly discarding clips that exhibited low information density or lacked clear audio-visual demonstrations. For the valid sessions, annotators selected high-value danmaku and precisely localized the response intervals $[t_{\text{start}}, t_{\text{end}}]$ grounded in specific audio-visual segments. To preserve the multi-party nature of live streaming, high-value yet unanswered queries were also retained and assigned to corresponding dialogue turns. Annotators further categorized the evidence source (Audio or Multimodal) and provided concise summaries for each turn. Finally, we conducted daily expert reviews on 10% of the samples to ensure consistency. The final dataset comprises ***8,053 video sessions*** and ***80,037 dialogue turns***, including 53.3K text-based turns and 26.7K video-based turns. Fig. 3 illustrates the distribution of dialogue turns and users, with detailed statistics and annotation protocols provided in Appendix A.1 and A.3.

## 4 Live-Aid Benchmark Construction

To enable comprehensive live interaction assessment, we curated a representative test set from

| Capability | Task | Type | Example |
|---|---|---|---|
| Multimodal Understanding | Video_QA | MCQ(4) | *Q: Is the decoration on this vest a circle of pearls?*<br>*(A) Yes, circle of pearls.* *(B) Other material. (C) No pearls. (D) Partial pearls.*<br>**Dmk@*737s*:** "Is this a pearl circle?" ⟶ **Response Clip:** [810s, 817s] |
| | | Open | *Q: How to connect the snaps to use as a double sleeping bag?*<br>*GT: Fold into rectangle, snap bottom/sides, then leg area.*<br>**Dmk@*61s*:** "How to snap this for double?" ⟶ **Response Clip:** [128s, 183s] |
| | MM_Cons | True/False | *Q: Please judge whether the following danmaku matches the given video clip [Video]$^{518s\text{-}532s}$: "Are these pants fleece-lined?"* ***(A) True*** *(B) False*<br>**Dmk@*564s*:** "Are these pants fleece-lined?" ⟶ **Response Clip:** [518s, 532s] |
| | | MCQ(2/4) | *Q: Select the video clip that best matches the danmaku: "Is it thick-soled?"*<br>*(A) [Video]$^{369s\text{-}384s}$ (B) [Video]$^{100s\text{-}126s}$ (C) [Video]$^{452s\text{-}457s}$* ***(D) [Video]$^{125s\text{-}159s}$***<br>**Dmk@*109s*:** "Is it thick-soled?" ⟶ **Response Clip:** [125s, 159s] |
| | Dmk_Intent | Class | *Based on History, Predict Next Dmk Intent. History: U:"Can kids eat this?" H:[Clip] ...**(6 Turns)** GT: Transaction Service - Price/Discount; Promo/Gift*<br>**Dmk@*386s*:** "Any promo gifts..." ⟶ **Response Clip:** [392s, 396s], ...**(6 Turns)** |
| Dialogue Modeling | Dmk_Gen | Open | *Q: Based on History & Curr Clip, Gen Dmk. History: U:"What size is the model..." H:[Clip] ...**(2 Turns)** [Curr Clip] GT: Are the eyeglasses the same?*<br>**Dmk@*108s*:** "What size is the model..." ⟶ **Response Clip:** [112s, 123s], ...**(2 Turns)** |
| | Dial_Summ | Open | *Q: Summarize the Dialogue: U:"Are sleeves long?" H:[Clip] ...**(6 Turns)***<br>*GT: Users ask about fit... Host confirms sleeves fit, ... length is 52cm.*<br>**Dmk@*243s*:** "Are sleeves long?" ⟶ **Response Clip:** [215s, 221s], ...**(6 Turns)** |
| Temporal Reasoning | Dmk_Order | Ranking | *Q: Sort shuffled danmaku based on ordered clips: [Clips]. Dmk: [0] "Is this padded?" [1] "Can you show the colors?" [2] "Is this in stock?" GT: [1 0 2]*<br>**Dmk@*31s*:** "Can you show the colors?" ⟶ **Response Clip:** [31s, 55s], ...**(4 Turns)** |
| | Timeline | Ranking | *Q: Sort shuffled clips based on ordered Dmk: [0] "How long is it?" [1] "Is the thick hoodie available?" [2] "Red hoodie collar?" [Shuffled Clips] GT: [3 1 0 2]*<br>**Dmk@*97s*:** "How long is it?" ⟶ **Response Clip:** [97s, 116s], ...**(4 Turns)** |
| | Evidence | Grounding | *Q: Input: Video (Dur: 532s), Frames: <N>, Timeline: <T>. Localize evidence for Dmk(@244s): "Is the outside cotton or fleece?" GT: [217s, 231s]*<br>**Dmk@*244s*:** "Is the outside cotton..." ⟶ **Response Clip:** [217s, 231s] **(History)** |

Table 2: **Task categories in Live-Aid.** We detail tasks and representative examples in Live Stream Interaction. *MCQ(2/4)* denotes Multiple-Choice with 2 or 4 options, *Class* denotes classification, and *Open* denotes Open-ended.

the Live-Aid corpus, comprising 1,646 clips and 16,090 dialogue turns with meticulous human annotations. Building on this foundation, we developed a Multi-Agent Collaborative Generation Pipeline inspired by recent agent-enhanced benchmarks (Wang et al., 2025c; Jiang et al., 2025).

## 4.1 Task Definitions

We categorize streaming dynamics into three temporal perspectives (*Instant*, *Delay*, and *History*) and abstract three core capabilities of live interactions (Chen et al., 2024a), as illustrated in Tab. 2. Specifically, **Multimodal Understanding (MU)** targets dynamic visual comprehension and user intent recognition (e.g., Video QA, Multi-Modal Consistency, Danmaku Intent Prediction); **Dialogue Modeling (DM)** assesses generative proficiency in simulation and summarization (e.g., Contextual Danmaku Generation, Structured Dialogue Summarization); and **Temporal Reasoning (TR)** evaluates timeline navigation and reasoning over temporal structures (e.g., Danmaku Ordering, Clip Timeline Reconstruction, Video Evidence Grounding). Detailed task definitions are in Appendix B.1.

## 4.2 Agent-Enhanced Benchmark

We implement a Multi-Agent Collaborative Generation Framework, consisting of four specialized agent groups, to construct intricate temporal reasoning and fine-grained perception tasks for *Live-Aid*, as illustrated in Fig. 2. Detailed prompts and workflows are provided in Appendix B.2.

**Modality Enrichment Agents.** As the perception foundation, the Modality Enrichment Agent aligns raw human annotations with video signals. We employ *rule-based agents* for precise timestamp synchronization and metadata extraction (e.g., duration, product info). Furthermore, to compensate for visual details overlooked in human annotations, we utilize Qwen3-VL (Bai et al., 2025a) to generate dense, time-stamped captions, capturing scene changes and OCR text to enrich the visual context.

**Component Generation Agents.** Acting as the pre-processing engine, this group constructs semantic building blocks and difficulty-controlling elements. We leverage GPT-4.1 (OpenAI, 2025b)

| Model | Size | Frames | VQA | MMC | DIP | CTR | | $DO_{Clip}$ | | $DO_{Full}$ | | VEG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Acc | Acc | Acc | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | tIoU |
| *Heuristic baselines* | | | | | | | | | | | | |
| Random Choice | - | - | 29.4 | 38.5 | 5.6 | 2.4 | 2.3 | 1.1 | 1.3 | 4.3 | 4.9 | 3.0 |
| Human | - | - | 93.75 | 100.0 | 54.55 | 99.3 | 99.6 | 74.6 | 72.6 | 74.6 | 72.6 | 43.4 |
| *Open-source Video-LLMs* | | | | | | | | | | | | |
| Qwen3-VL-Thinking | 4B | 1fps† | 44.2 | 54.4 | 14.8 | N/A | N/A | -8.9 | -6.2 | 8.3 | 12.1 | 3.4 |
| | 8B | | 42.8 | 54.4 | 13.3 | N/A | N/A | -41.2 | -41.4 | 5.1 | 10.0 | 3.8 |
| Qwen3-VL-Instruct | 4B | 1fps† | 45.4 | 53.3 | 17.6 | -8.5 | -9.0 | 8.6 | 12.6 | 7.8 | 11.5 | 4.0 |
| | 8B | | 43.7 | 56.5 | 15.6 | -9.4 | -10.9 | 3.1 | 3.1 | 4.3 | 5.0 | 3.8 |
| InternVL3.5 | 4B | 32 | 45.0 | 46.1 | 15.6 | -0.4 | 0.3 | -7.2 | -6.4 | -4.8 | -4.2 | **5.5** |
| | 8B | | 44.6 | 52.2 | 14.7 | 0.4 | 1.2 | -2.4 | -1.0 | 0.7 | 2.2 | 2.4 |
| Keye-VL-1.5 | 8B | 1fps† | 51.3 | 52.4 | 12.6 | -16.2 | -20.2 | -4.7 | -4.7 | -3.6 | -4.3 | 5.1 |
| VideoLLaMA3 | 7B | 1fps† | 43.5 | 49.0 | 11.7 | -3.5 | -6.2 | -6.3 | -7.3 | -5.0 | -6.1 | 5.4 |
| MiniCPM-V 4.5 | 8B | 1fps† | 50.7 | 54.5 | 19.1 | 6.4 | 8.2 | **10.7** | **15.0** | 5.4 | 7.2 | 3.4 |
| LLaVA-OneVision | 7B | 32 | 40.7 | 45.2 | 19.0 | -6.1 | -8.4 | 7.3 | 11.7 | 9.9 | **16.1** | 2.9 |
| LLaVA-OneVision-1.5 | 8B | 1fps† | 40.5 | 52.2 | **21.9** | -0.2 | -0.5 | 6.8 | 10.2 | **10.4** | 15.7 | 4.8 |
| *Open-source Omni-LLMs* | | | | | | | | | | | | |
| VideoLLaMA2.1-AV | 7B | 32 | 38.5 | 59.1 | 8.6 | -21.7 | -35.8 | 7.6 | 12.3 | 3.8 | 8.2 | 4.1 |
| +audio | | | 39.2 | 58.2 | 8.6 | -19.3 | -32.2 | 8.8 | 14.4 | 6.1 | 11.1 | 4.2 |
| Qwen3-Omni-30B | A3B | 1fps† | 53.3 | **61.5** | 16.9 | -9.8 | -9.6 | -1.6 | -1.9 | -5.6 | -6.5 | 2.4 |
| +audio | | | 59.1 | 61.4 | 16.7 | -7.2 | -5.8 | -2.8 | -3.8 | -6.2 | -7.1 | 2.6 |
| OmniVinci | 9B | 32 | 52.6 | 54.0 | 20.6 | -6.0 | -11.4 | -8.5 | -11.2 | -3.4 | -2.7 | 1.9 |
| +audio | | | **83.8** | 57.3 | 17.6 | -5.2 | -10.2 | -6.8 | -8.7 | -5.3 | -5.7 | 2.9 |
| *Commercial MLLMs* | | | | | | | | | | | | |
| GPT-4o-mini | ~8B | 1fps† | 27.3 | 48.7 | 16.6 | -6.0 | -7.8 | 7.6 | 7.9 | 8.0 | 8.6 | 2.3 |
| GPT-4o | ~200B | | 35.3 | 56.8 | 12.0 | **18.0** | **20.9** | 5.4 | 6.1 | -43.6 | -43.0 | 3.7 |
| GPT-4.1 | - | | 37.0 | 60.5 | 17.9 | 7.1 | 8.1 | 7.7 | 8.5 | 3.9 | 5.0 | 4.7 |

Table 3: **Performance of MLLMs on LiveAid across MU and TR.** Size means the LLM size. Reported metrics are Accuracy, Kendall's Tau ($\tau$), Spearman's Rank Correlation ($\rho$), and tIoU (all in %). † Videos are sampled at 1 fps with an upper frame limit. The best, second-best, and third-best results are marked purple, orange, and gray, respectively. N/A indicates no valid answer within default max thinking length.

to refine colloquial human summaries into structured abstracts and dynamically evolve user intent categories. Furthermore, the *Adversarial Distractor Agent* generates challenging distractors via a stratified sampling strategy, comprising both inter-video *Global Negatives* and harder intra-video *Local Negatives* to ensure robust discrimination.

**Task Synthesis Agents.** Leveraging enriched multimodal contexts and pre-computed components, this group utilizes GPT-4.1 to synthesize final tasks across MU, DM, and TR dimensions. By integrating dialogue history, danmaku, visual evidence, and adversarial distractors, it assembles standardized task instances (e.g., multiple-choice QA, True/False, ranking, and temporal grounding) that rigorously align with our task definitions.

**Verification Agents.** To guarantee the solvability and correctness of the generated samples, we adopt a Consensus Committee mechanism in which GPT-4.1, Qwen3, and Qwen3-VL independently solve each task. Only samples where the committee reaches a consensus on the answer are retained. We then conduct a rigorous human review to eliminate subtle ambiguities and correct domain-specific cultural or temporal misinterpretations, thereby establishing the high-fidelity *Live-Aid* benchmark.

### 4.3 Comparison with Previous Work

Previous dialogue datasets primarily treat video as a static visual context for text-only turns, often overlooking the interleaved dynamics of live streaming and lacking precise temporal annotations. Moreover, existing datasets predominantly focus on event-centric Video QA, largely ignoring the temporal reasoning and generative interactions essential for live streams. In contrast, as shown in Tab. 1, our **Live-Aid** captures high-density viewer-host interactions grounded in vivid audio-visual evidence, making it the first large-scale interleaved live dialogue dataset with fine-grained temporal annotations. Furthermore, by leveraging high-quality human annotations within a multi-agent collaborative framework, Live-Aid offers a comprehensive evaluation suite for interleaved live interactions.

## 5 Experiment

### 5.1 Experimental Setups

**Metrics.** We evaluate performance using two approaches (Lei et al., 2025): **(1)** Automatic Metrics: We use accuracy for selection tasks, Kendall's $\tau$ and Spearman's $\rho$ for ordering, and tIoU for temporal grounding. Open-ended responses are assessed

| Model | Size | Frames | VQA BL-2 | VQA MET | VQA $F1_{BERT}$ | VQA $S_{GPT4.1}$ | SDS BL-2 | SDS MET | SDS $F1_{BERT}$ | SDS $S_{GPT4.1}$ | CDG BL-2 | CDG MET | CDG $F1_{BERT}$ | CDG $S_{GPT4.1}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Open-source Video-LLMs* | | | | | | | | | | | | | | |
| Qwen3-VL-Thinking | 4B | 1fps† | 13.12 | 26.85 | 67.51 | 46.69 | 18.71 | 30.01 | 73.75 | 58.40 | 0.67 | 4.15 | 53.55 | 62.25 |
| | 8B | | 8.63 | 20.96 | 62.82 | 41.38 | 15.77 | 27.48 | 72.20 | 58.64 | 1.45 | 5.89 | 56.23 | 70.26 |
| Qwen3-VL-Instruct | 4B | 1fps† | 9.8 | 18.68 | 65.16 | 40.54 | **19.22** | **31.54** | **74.56** | 56.92 | 1.27 | 4.42 | 56.36 | 64.97 |
| | 8B | | 9.98 | 20.00 | 65.07 | 38.69 | 14.46 | 28.75 | 71.90 | 58.10 | 1.99 | 6.02 | 57.78 | 69.02 |
| InternVL3.5 | 4B | 32 | 2.72 | 6.78 | 57.48 | 35.31 | 10.86 | 22.62 | 69.58 | 51.74 | 1.13 | 5.14 | 54.79 | 55.89 |
| | 8B | | 2.75 | 6.59 | 56.53 | 36.59 | 9.92 | 23.63 | 69.29 | 53.66 | 1.07 | 5.09 | 54.86 | 56.60 |
| Keye-VL-1.5 | 8B | 1fps† | 3.59 | 7.81 | 56.04 | 40.08 | 10.81 | 23.04 | 68.60 | 53.47 | 2.30 | 4.68 | 59.35 | 65.78 |
| VideoLLaMA3 | 7B | 1fps† | 7.32 | 15.33 | 64.09 | 41.31 | 14.11 | 22.60 | 71.14 | 52.14 | 1.56 | 4.15 | 57.12 | 63.27 |
| MiniCPM-V 4.5 | 8B | 1fps† | 5.16 | 10.59 | 59.63 | 40.62 | 11.52 | 26.56 | 70.82 | 55.46 | 1.01 | 4.90 | 55.49 | 65.43 |
| LLaVA-OneVision | 7B | 32 | 9.3 | 16.7 | 63.47 | 40.77 | 5.40 | 14.47 | 64.60 | 41.15 | 1.73 | 4.78 | 56.34 | 49.03 |
| LLaVA-OneVision-1.5 | 8B | 1fps† | 10.32 | 24.73 | 65.18 | 45.81 | 15.66 | 24.23 | 72.56 | 48.22 | 1.07 | 4.84 | 54.64 | 62.84 |
| *Open-source Omni-LLMs* | | | | | | | | | | | | | | |
| VideoLLaMA2.1-AV | 7B | 32 | 9.42 | 18.18 | 61.92 | 40.23 | 6.77 | 14.2 | 64.98 | 41.52 | 1.80 | 6.33 | 56.56 | 53.07 |
| +audio | | | 7.29 | 15.44 | 60.18 | 41.23 | 5.84 | 12.66 | 64.32 | 39.84 | 1.79 | 5.46 | 56.22 | 51.18 |
| Qwen3-Omni-30B | A3B | 1fps† | 8.89 | 21.60 | 64.35 | 39.31 | 18.09 | 30.94 | 73.68 | 59.41 | 3.06 | **10.11** | 59.37 | 72.56 |
| +audio | | | 8.46 | 21.59 | 64.60 | 39.77 | 17.76 | 30.31 | 73.44 | 59.01 | **3.16** | 10.10 | **59.38** | 73.10 |
| OmniVinci | 9B | 32 | 10.61 | 19.88 | 65.34 | 44.23 | 14.88 | 23.64 | 71.24 | 49.73 | 1.65 | 6.24 | 56.64 | 61.12 |
| +audio | | | **16.45** | 26.93 | 69.28 | **54.77** | 15.86 | 25.70 | 71.66 | 54.99 | 2.26 | 8.05 | 58.02 | 66.77 |
| *Commercial MLLMs* | | | | | | | | | | | | | | |
| GPT-4o-mini | ~8B | 1fps† | 14.49 | 26.14 | 68.29 | 39.73 | 15.17 | 27.36 | 72.63 | **60.23** | 2.14 | 7.78 | 57.61 | 70.50 |
| GPT-4o | ~200B | | 7.45 | 13.51 | 59.43 | 29.08 | 13.02 | 25.80 | 71.71 | 58.23 | 2.68 | 9.15 | 58.87 | **73.15** |
| GPT-4.1 | - | | 15.00 | **28.12** | **69.87** | 47.38 | 9.12 | 28.28 | 69.92 | 59.41 | 2.57 | 9.79 | 58.62 | 73.04 |

Table 4: **Performance of MLLMs on LiveAid across open-ended question types.** Reported metrics are BLEU-2(**BL-2**), METEOR (**MET**), BERTScore F1 (**$F1_{BERT}$**), and GPT-4o scores (**$S_{GPT4.1}$**) (all in %).

via BLEU-2, METEOR, and BERTScore (F1). **(2)** LLM-as-a-Judge: We employ GPT-4.1 to evaluate response quality on a scale of 0 to 10.

**Models.** We conduct comprehensive experiments on Live-Aid using a diverse set of state-of-the-art open-source and closed-source Video-MLLMs: (a) Open-source Video-LLMs, including Qwen3-VL (Bai et al., 2025a), InternVL3.5 (Wang et al., 2025a), Keye-VL-1.5 (Yang et al., 2025b), VideoLLaMA3 (Zhang et al., 2025), LLaVA-OneVision-1.5 (An et al., 2025), and MiniCPM-V-4.5 (Yu et al., 2025). (b) Open-source Omni-LLMs with native audio perception: VideoLLaMA2.1-AV (Cheng et al., 2024), Qwen3-Omni (Xu et al., 2025), and OmniVinci (Ye et al., 2025). (c) Proprietary MLLMs, such as GPT-4o (Hurst et al., 2024) and GPT-4.1 (OpenAI, 2025b). Detailed inference configurations (such as frame sampling rates and multi-modal input formats) are provided in Appendix C.

### 5.2 Results and Analysis

We employ random selection as a heuristic baseline, presenting results for discriminative tasks in Tab. 3, generative tasks in Tab. 4, and fine-grained performance on Danmaku Intent Prediction in Tab. 5. We further investigate the comprehensive capabilities of MLLMs in real-world live streaming scenarios through three guiding research questions (RQs): (1) the capability in handling interleaved live interaction; (2) the contribution of audio perception; and (3) the impact of dialogue context length on viewer-centric danmaku generation.

**RQ1: How do MLLMs perform when facing complex interleaved live interactions?** While MLLMs demonstrate promising proficiency in basic video understanding tasks, their performance declines sharply as inputs shift toward complex interleaved interactions, especially for reasoning-oriented models (e.g., Qwen3-VL-Thinking) which often struggle to complete responses under context constraints. Meanwhile, Omni-LLMs (e.g., OmniVinci) leverage comprehensive audio-visual perception to significantly outperform standard Video-LLMs in Video QA. In generative tasks (e.g., danmaku generation), MLLMs often underperform due to the high diversity of user tones and phrasing styles inherent in live scenarios. Furthermore, temporal reasoning remains a significant challenge, particularly in evidence grounding, where sparse frame sampling limits the ability to precisely locate interactions within long-duration streams.

**RQ2: To what extent does audio perception contribute in live streaming scenarios?** Ablation studies on Omni-LLMs show that incorporating audio input consistently improves performance over vision-only settings, underscoring the importance of audio perception for understanding high-density live-streaming dialogues. Among them, VideoLLaMA-2.1-AV demonstrates limited audio proficiency, significantly lagging behind OmniVinci and Qwen3-Omni. However, despite the overall benefits of audio modalities, performance remains limited on interleaved video–text tasks, suggesting insufficient multimodal integration.

**RQ3: How does the dialogue context length impact viewer-centric danmaku generation?** We investigate the impact of context length by varying the number of preceding dialogue turns, as illustrated in Fig. 4. Generally, performance on both tasks improves as more historical turns are incorporated, indicating that longer contexts help models capture the temporal evolution of user intent and thus lead to improved performance in both prediction and generation. However, for the Danmaku Intent Prediction task, performance begins to degrade when the history exceeds 10 turns. We attribute this drop to a trade-off between context length and visual density. Under a fixed total frame budget, extending the dialogue context necessarily reduces the number of frames allocated to each individual turn, resulting in insufficient visual evidence for accurate intent inference. Additional analysis using a longer dialogue context is provided in Appendix D.

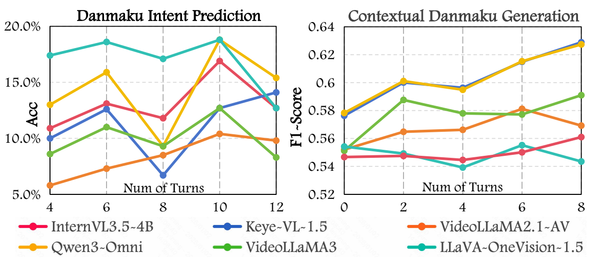



Figure 4: **Effect of dialogue context length.**

| Model | OA | DIP | | |
|---|---|---|---|---|
| | | PI | TS | CC |
| Qwen3-VL$_{4B}$ | 16.2 | 13.8 | 12.5 | 0.0 |
| Qwen3-VL$_{8B}$ | 14.5 | 13.2 | 8.8 | 1.1 |
| InternVL3.5$_{4B}$ | 15.6 | 15.2 | 7.3 | 0.0 |
| InternVL3.5$_{8B}$ | 14.7 | 14.5 | 5.7 | 4.1 |
| Keye-VL-1.5 | 12.6 | 7.0 | 18.0 | 0.0 |
| VideoLLaMA3 | 11.7 | 10.5 | 6.5 | 3.8 |
| MiniCPM-V 4.5 | 19.1 | 18.4 | 8.7 | 6.1 |
| LLaVA-OneVision | 19.0 | 20.3 | 4.6 | 2.2 |
| LLaVA-OneVision-1.5 | 21.9 | 24.2 | 5.2 | 0.0 |
| VideoLLaMA2.1-AV | 8.6 | 7.8 | 4.4 | 1.9 |
| Qwen3-Omni-30B | 16.7 | 17.3 | 5.9 | 0.0 |
| OmniVinci | 17.6 | 16.3 | 10.9 | 0.0 |
| GPT-4o-mini | 16.6 | 14.3 | 8.8 | 17.0 |
| GPT-4o | 12.0 | 7.8 | 12.7 | 8.9 |
| GPT-4.1 | 17.9 | 18.4 | 6.2 | 0.0 |

Table 5: **Fine-grained performance on the DIP Task.** PI: Product Information; TS: Transaction Service; CC: Chitchat. Results are reported as Accuracy (%).

### 5.3 Detailed Analysis on Temporal Dynamics

To assess temporal grounding capabilities, we evaluate models on three query types: *Instant* (evidence immediately follows the query), *Delay* (evidence appears after a temporal gap), and *History* (evidence precedes the query, a particularly challenging setting that requires retrieval against the temporal flow). As illustrated in Fig. 5, MLLMs exhibit substantially stronger performance on the *Instant* type. This trend suggests an implicit bias toward immediate chronological succession, where models assume relevant evidence occurs in close temporal proximity following the query. In contrast, performance degrades markedly for the *Delay* type, indicating difficulties in modeling dependencies over extended temporal spans. Most notably, performance on the *History* type is extremely limited, with an average tIoU of only ∼1%, highlighting a fundamental deficiency in reasoning about reverse temporal causality. Detailed results for all models are provided in Appendix D.

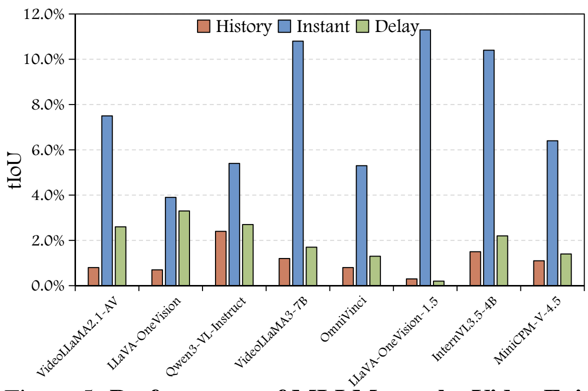



Figure 5: **Performance of MLLMs on the Video Evidence Grounding task across three temporal types.** Results are reported as tIoU (%).

## 6 Conclusion

The advancement of MLLMs in live scenarios has been constrained by the scarcity of high-quality data capturing dynamic, interleaved interactions. To bridge this gap, we introduce **Live-Aid**, a unified resource comprising the first large-scale live dialogue dataset with meticulous temporal alignments and a comprehensive benchmark. By leveraging a multi-agent collaborative framework, we tailor eight core evaluation tasks across three key dimensions to assess MLLMs in live streaming scenarios. Extensive experiments demonstrate that state-of-the-art MLLMs continue to struggle with the high-density, multi-turn dynamics of live streams, particularly in scenarios requiring fine-grained temporal grounding and audio-visual synchronization. We hope our **Live-Aid** will serve as foundational resources for research on authentic live streaming interaction and live agents.

## Limitations

Despite the extensive contributions of **Live-Aid**, we acknowledge limitations regarding computational efficiency and resource demands. Evaluating MLLMs on long-context videos is inherently computationally intensive, as current state-of-the-art models require processing extensive sequences of high-density interleaved data. Consequently, these substantial memory and inference requirements may pose challenges for efficient experimentation and restrict the dataset's accessibility for researchers with constrained resources. In addition, **Live-Aid** is limited to Chinese e-commerce live streams from a single platform, which may limit generalization to other languages or domains. However, the dataset focuses on core capabilities of live-streaming interaction, suggesting potential generalizability beyond the current domain.

## Ethics Statement

The construction and development of *Live-Aid* involve publicly accessible live streaming videos and real-time user interactions (*danmaku*). We address ethical considerations related to data privacy, content safety, annotator welfare, and responsible use.

**Data Privacy.** All data were collected in compliance with the terms of service of the source platforms and applicable data protection regulations. The videos consist of public live streams in e-commerce scenarios involving public-facing hosts. User-generated *danmaku* were processed through a strict anonymization pipeline that removes all Personally Identifiable Information (PII), including user identifiers and metadata, ensuring that individual viewers cannot be identified.

**Content Safety.** Given the unscripted nature of live streaming, we apply a hybrid filtering pipeline combining automated screening and expert review to exclude hate speech, explicit content, and harassment. We note that e-commerce streams may exhibit commercial biases and encourage cautious interpretation of evaluation results.

**Annotator Welfare.** Ten expert annotators were employed for temporal grounding and verification tasks. All annotators were fairly compensated above local minimum wage, and the annotation process was designed to limit cognitive load through reasonable task allocation and regular breaks.

**Responsible Use.** *Live-Aid* is intended to advance research on multimodal understanding of temporally structured human interactions. We discourage misuse of the dataset for deceptive marketing, sentiment manipulation, or spam generation, and advocate its use for improving user experience, accessibility, and transparent human–AI interaction.

## Acknowledgements

This work is supported by "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01020), the National Natural Science Foundation of China (No. 62406015), research funding from Kuaishou Technology, the Emergency Management Research and Development Project of Zhejiang Province (No. 2024YJ018).

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Aihui Chen, Yaning Chen, Ruohan Li, and Yaobin Lu. 2024a. The interaction between the anchor and customers in live-streaming e-commerce. *Industrial Management & Data Systems*, 124(6):2151–2179.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024b. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15387–15405.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. 2025. Online video understanding: Ovbench and videochat-online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3328–3338.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Wiradee Imrattanatrai, Masaki Asada, Kimihiro Hasegawa, Zhi-Qi Cheng, Ken Fukuda, and Teruko Mitamura. 2025. A video-grounded dialogue dataset and metric for event-driven activities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24203–24211.

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

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2024. Dialogcc: An automated pipeline for creating high-quality multimodal dialogue dataset. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1938–1963.

Yiming Lei, Chenkai Zhang, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. 2025. GODBench: A benchmark for multimodal large language models in video comment art. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11884–11952, Vienna, Austria. Association for Computational Linguistics.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025a. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xingyu Li, Chen Gong, and Guohong Fu. 2025b. Multimodal coreference resolution for Chinese social media dialogues: Dataset and benchmark approach. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31513–31525, Vienna, Austria. Association for Computational Linguistics.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024b. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, and 1 others. 2023. Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1303–1313.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1024–1034.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. 2024. Mmlscu: A dataset for multi-modal multi-domain live streaming comment understanding. In *Proceedings of the ACM Web Conference 2024*, pages 4395–4406.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, and 1 others. 2025. Ovobench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18902–18913.

OpenAI. 2025a. Gpt-5 system card. https://openai.com/zh-Hans-CN/index/gpt-5-system-card/. Accessed: 2025-12-28.

OpenAI. 2025b. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-12-28.

Chenhui Qiang, Zhaoyang Wei, Xumeng Han, Zipeng Wang, Siyao Li, Xiangyuan Lan, Jianbin Jiao, and Zhenjun Han. 2025. Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12698–12705.

Xiaoming Shi, Zeming Liu, Yiming Lei, Chenkai Zhang, Haitao Leng, Chuan Wang, Qingjie Liu, Wanxiang Che, and Yunhong Wang. 2025. Kwaichat: A large-scale video-driven multilingual mixed-type dialogue corpus. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2279–2294.

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2025. Audio-visual llm for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4246–4255.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2020. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429.

Yuchen Song, Andong Chen, Wenxin Zhu, Kehai Chen, Xuefeng Bai, Muyun Yang, and Tiejun Zhao. 2025. Culture in a frame: C3b as a comic-based benchmark for multimodal culturally awareness. *arXiv preprint arXiv:2510.00041*.

Haoran Sun, Haoyu Bian, Shaoning Zeng, Yunbo Rao, Xu Xu, Lin Mei, and Jianping Gou. 2025. Datasetagent: A novel multi-agent system for auto-constructing datasets from real-world images. *arXiv preprint arXiv:2507.08648*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Qun Liu, and Dongyan Zhao. 2025b. Friends-mmc: A dataset for multi-modal multi-party conversation understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25425–25433.

Zihan Wang, Xiaocui Yang, Yongkang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025c. Muse: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1027–1053.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting

He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025b. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6142–6152.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. Mmchat: Multi-modal chat dataset on social media. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 5778–5786.

# A More Details of Dialogue Construction

## A.1 Overview of Live-Aid

We present the comprehensive statistics of the human-annotated **Live-Aid** in Tab. 6, organized into *Live Information*, *Dialogue Configuration*, and *Audience Engagement*. As shown in Tab. 6, Live-Aid is characterized by its large-scale and high-density nature, spanning over **1,100 hours** derived from 1,763 real-world source streams and encompassing 44 distinct product categories (the distribution of the categories is illustrated in Fig. 6). Regarding *Dialogue Configuration*, the dataset features **80,037 interleaved turns**, where viewer Danmaku is temporally aligned with the streamer's video responses, addressing the scarcity of high-quality interleaved data in existing research. Furthermore, the *Audience Engagement* statistics involving **46,819 unique users** underscore the dynamic and evolving context of the dataset, distinguishing Live-Aid from traditional static or narrative-driven video benchmarks. While the current dataset is grounded in Chinese e-commerce live streams, its task design targets domain-agnostic core capabilities of live streaming, allowing it to generalize to other scenarios such as education, gaming, entertainment, and social streaming. Moreover, it may support future research on culturally aware live interaction, such as cross-cultural understanding and culturally grounded response generation (Song et al., 2025).

## A.2 Data Quality Control Processing

To construct a high-quality benchmark from noisy raw live streams, we implemented a comprehensive three-stage cleaning pipeline. We detail the specific steps and criteria below.

**Session Segmentation.** Raw live streams typically span several hours and often contain multiple product introductions. To segment these continuous streams into coherent semantic units, we employed a robust multimodal approach. Specifically, we used Automatic Speech Recognition (ASR) to detect pauses longer than 1.5 seconds of silence and combined this information with visual scene detection—based on color histogram differences exceeding 0.6—to reliably identify product transitions.

| Live Information | |
|---|---|
| Total Source Live Streams | 1,763 |
| Total Video Sessions | 8,053 |
| Product Categories | 44 |
| Total Video Duration | 1,102.2 hour |
| - Avg. Session Duration | 492.7 second |
| **Dialogue Configuration** | |
| Total Dialogue Turns | 80,037 |
| - Total Text Turns | 53,319 |
| - Total Video Turns | 26,718 |
| - Avg. Turns per Session | 9.94 |
| **Audience Engagement** | |
| Total Danmaku Messages | 53,319 |
| Total Unique Users | 46,819 |
| - Avg. Danmaku per Session | 6.62 |
| - Avg. Users per Session | 5.81 |

Table 6: **Statistics of the Live-Aid Dialogue.** *Unique Users* refers to distinct individuals after deduplication.

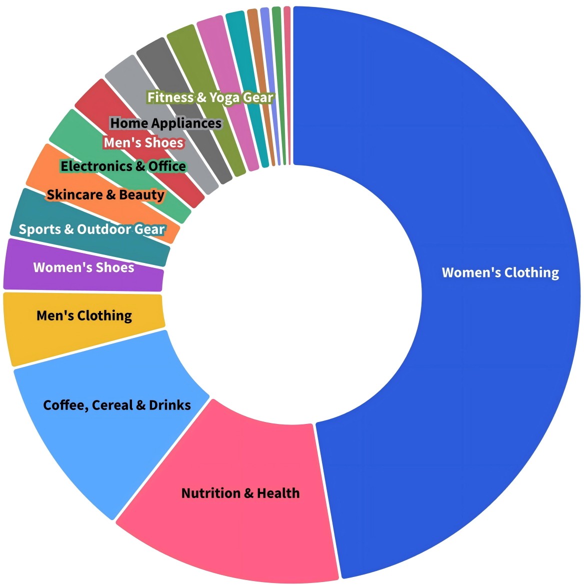



Figure 6: Distribution of the 44 product categories in the Live-Aid benchmark.

This process yielded an initial set of 308.9K candidate video segments.

**Danmaku Filtering.** Real-time comments in live streams are inherently noisy. To ensure a high signal-to-noise ratio, we applied a hierarchical filtering strategy comprising three distinct mechanisms. First, we employed Rule-based Heuristics via Regular Expressions (Regex) to eliminate low-quality noise, including repetitive patterns (e.g., "hahaha" or numeric slang), platform-generated system notifications, and short texts (fewer than 2 characters or emojis-only). Second, we implemented a Blacklist Mechanism based on common toxic lexicons to filter out profanity, advertisements, and sensitive political topics. Finally, we utilized Intent-based Semantic Filtering to distinguish meaningful dialogue from trivial chatter. We leveraged Qwen3-30B to classify danmaku intents into *Information-Seeking* (e.g., inquiries about size or price) versus *Chit-Chat*, retaining only those identified as genuine information-seeking queries.

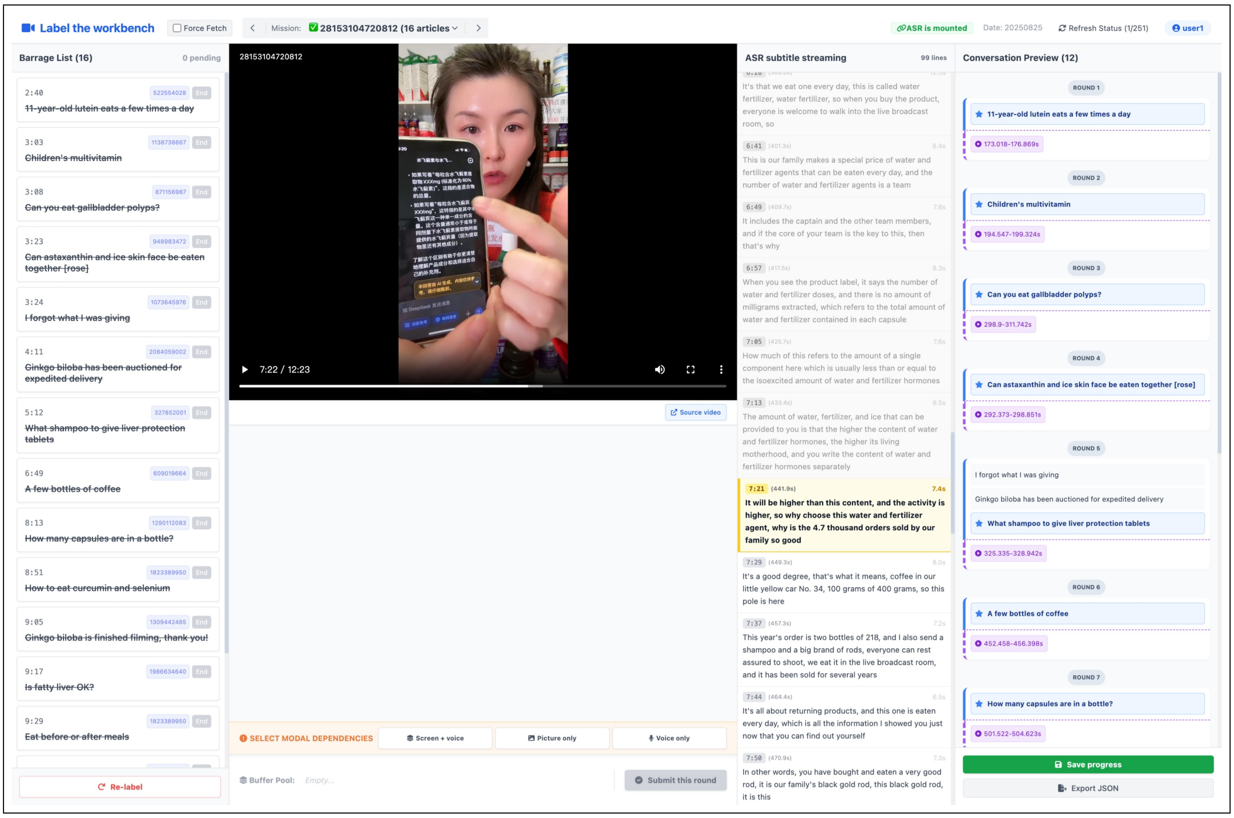
**Session Filtering.** To ensure the data is suitable for multimodal modeling, we filtered out segments with low audio quality or sparse user interactions. We imposed duration constraints ($[1, 20]$ minutes) to fit the context window limits of current Video-LLMs. Furthermore, to guarantee a realistic multi-party interaction setting, we restricted the number of unique users per session to the range of $[4, 30]$. This rigorous processing stage resulted in a candidate pool of 20,406 high-quality sessions, which were subsequently submitted for human annotation.

### A.3 Human Annotation Details

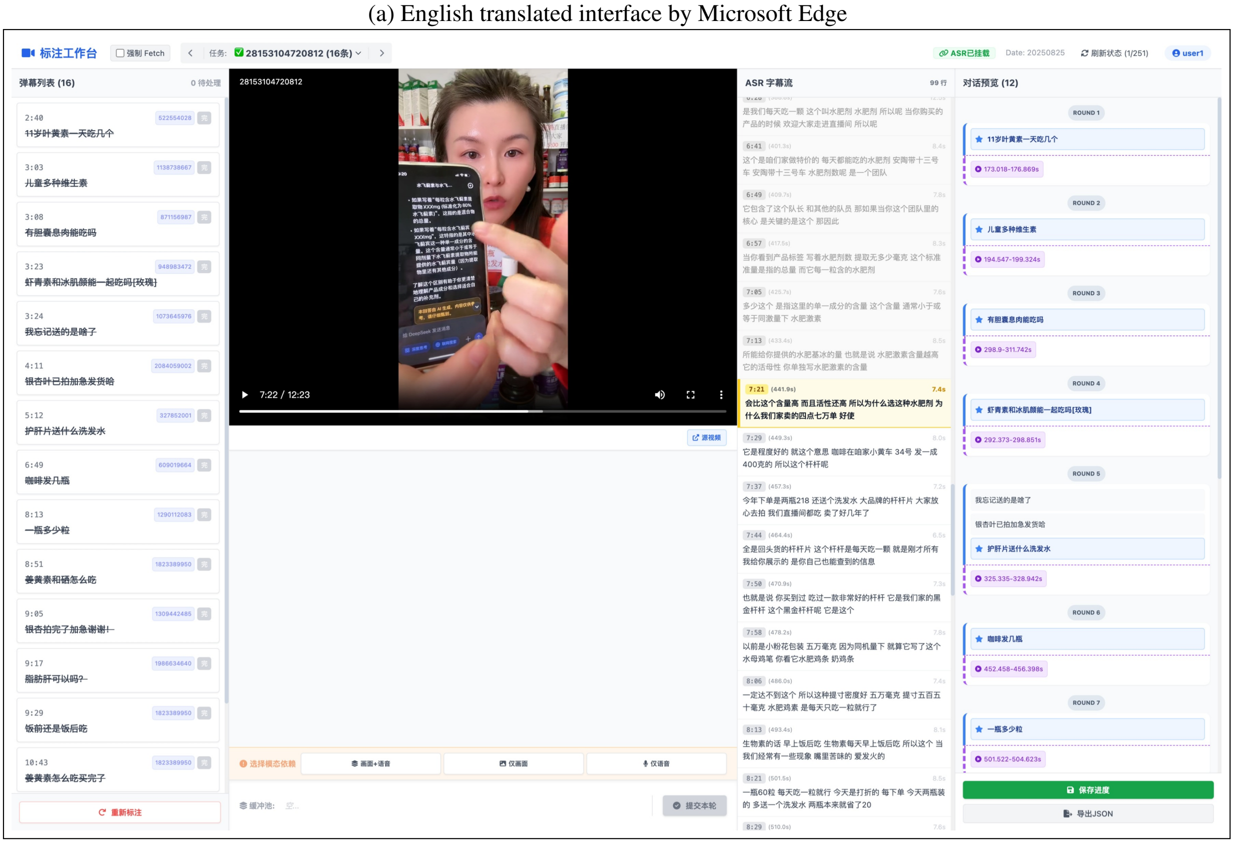
We developed a customized annotation platform and trained a team of 10 professional workers to perform the annotation tasks. As illustrated in Fig. 7, the platform adopts an intuitive three-panel design: the left panel lists the chronological raw danmaku; the center panel allows annotators to scrub the video and mark start/end timestamps on a multi-track timeline; and the right panel organizes the selected content into dialogue turns, facilitating the management of multimodal responses. Based on this platform, the annotation pipeline consists of four specific phases.

**1) Quality Filtering and Temporal Grounding.** As a prerequisite, annotators served as a final quality filter, manually discarding clips that exhibited low information density or lacked clear audio-visual demonstrations. For the valid sessions, annotators scanned the danmaku stream to identify the **Target Danmaku** ($d^*$)—the specific query that triggers the streamer's response. Once identified, they explicitly localized the precise timestamp interval $[t_{\text{start}}, t_{\text{end}}]$ grounded in specific audio-visual segments. The protocol defined the start time as the moment the streamer begins to address the topic (verbally or by interaction) and the end time as the completion of the explanation.

**2) Multi-party Grouping.** To faithfully reconstruct the multi-party interaction field, we did not discard high-value queries that were ignored by the host. Instead, annotators were instructed to assign




(a) English translated interface by Microsoft Edge




(b) Original Chinese Interface

Figure 7: **The Customized Annotation Platform.** The interface features a three-column layout designed for efficiency: the ***Left Panel*** displays the candidate danmaku stream; the ***Center Panel*** provides the video player with a precision timeline for frame-level temporal grounding; and the ***Right Panel*** serves as the dialogue construction workspace, where annotators define turns, drag-and-drop contextual danmaku, and verify ASR transcripts.

these unanswered danmaku to the specific dialogue turn ($T_k$) in which they appeared. These comments constitute the **Contextual Danmaku Pool ($\mathcal{C}_k$)**, forming the ambient context that the model must perceive but distinguish from the target intent.

**3) Modality and Summarization.** To support fine-grained diagnostics, annotators classified the **Video Evidence Source** required to answer the target query into two distinct categories: *Audio-only* (where the answer is purely verbal, e.g., stating the price or checking stock) and *Multimodal* (where the answer relies on visual demonstrations aligned with verbal explanations, e.g., showing the lining of a coat or trying on clothes). Additionally, a concise textual summary was written for each turn to facilitate high-level content understanding.

**4) Quality Assurance.** We implemented a two-stage verification process. First, the annotation platform deployed heuristic checks (e.g., ensuring $t_{end} > t_{start}$). Second, senior experts conducted daily reviews, randomly sampling 10% of the data. During this phase, any ambiguous cases were resolved through expert adjudication to maintain strict adherence to the annotation protocols, ensuring high consistency in both temporal localization and attribute labeling. This rigorous process yielded a final set of 8,053 annotated video sessions, comprising approximately 5.2k Audio-only instances and 2.9k Multimodal instances.

### A.4 Qualitative Examples

To provide a concrete visualization of the dataset quality and diversity, we present four representative annotated cases in Fig. 18-Fig. 21. These examples vividly illustrate the high-density interleaved interactions characteristic of live streaming. Specifically, they showcase the precise temporal alignment between viewer Danmaku and the streamer's dynamic audio-visual responses, presenting scenarios where accurate answers are unattainable via pure text alone, thus highlighting distinct evidence sources ranging from purely verbal explanations to interleaved multimodal interactions.

## B More Details of Benchmark Construction

### B.1 Detailed Task Definitions

Drawing insights from foundational works in visual dialogue and live interaction analysis (Chen et al., 2024a; Gao et al., 2023; Li et al., 2024b; Niu et al., 2025), we have formulated a structured evaluation framework tailored to the *interleaved live interaction* paradigm. Unlike generic video benchmarks, our task definitions are specifically designed to probe core capabilities required for processing high-density, dynamic multimodal streams, such as temporal grounding and intent recognition in evolving contexts. A comprehensive comparison with existing benchmarks is provided in Tab. 7, and the fine-grained statistics for each constructed task are summarized in Tab. 8. Below, we elaborate on the specific definitions and evaluation goals for the tasks within each capability dimension.

**Multimodal Understanding (MU).** This capability assesses the model's precision in perceiving dynamic visual content and aligning it with noisy textual streams. We define three specific tasks: (1) *Video QA* serves as a fundamental perception test, requiring the model to answer natural language questions based on visual cues (e.g., actions, objects) or spoken content. (2) *Multimodal Consistency* acts as a hallucination probe, determining whether a textual statement is factually supported by the video evidence. To evaluate robustness, we stratify this task into four difficulty levels: *Diff1 (Intra-video Distraction)* presents the correct answer alongside distractors sampled from the same video source; *Diff2 (Inter-video Distraction)* presents the correct answer with distractors from different videos; *Diff3 (Pure Negative)* requires rejecting all options ("None of the above") where all candidates are from external sources; and *Diff4 (Hard Rejection)* requires rejection despite the presence of decoys from the same video, serving as a strong inductive trap. (3) *Danmaku Intent Prediction* focuses on social understanding, where the model must classify the underlying user intent into 18 fine-grained categories spanning three domains: *Chit-chat* (e.g., action requests, casual talk), *Transaction Services* (e.g., price inquiries, logistics, after-sales), and *Product Information* (e.g., usage experience, specs, design).

**Dialogue Modeling (DM).** This capability evaluates proficiency in information generation and compression within the live interaction loop. The *Contextual Danmaku Generation* task simulates active participation, requiring the model to generate relevant, style-consistent danmaku based on the current video stream and dialogue history. In contrast, *Structured Dialogue Summarization* requires the model to act as an observer, compressing long sequences of interleaved streamer speech and audience comments into structured summaries that

| Benchmarks | Answer | M.I. | Source | Anno. | Capability | | | Video | | #QA Pairs |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | MU | DM | TR | #Clips. | Len(s) | |
| MVBench (Li et al., 2024a) | | ✗ | | | ✔ | ✗ | ✗ | 3,641 | 16.0 | 4,000 |
| MMBench-Video (Fang et al., 2024) | | ✗ | | | ✔ | ✗ | ✗ | 609 | 165.4 | 1,998 |
| AutoEval-Video (Chen et al., 2024b) | | ✗ | | | ✔ | ✗ | ✗ | 327 | 14.6 | 7,540 |
| VideoVista (Li et al., 2024b) | | ✗ | | | ✔ | ✗ | ✔ | 3,402 | 131.0 | 24,906 |
| Video-MME (Fu et al., 2025) | | ✗ | | | ✔ | ✗ | ✗ | 900 | 1017.9 | 2,700 |
| DanmakuTPP-QA (Jiang et al., 2025) | | ✗ | | | ✔ | ✗ | ✓ | 2,605 | 175.2 | 2,605 |
| OVBench (Huang et al., 2025) | | ✗ | | | ✔ | ✗ | ✓ | 1,463 | 151.7 | 7,090 |
| OVOBench (Niu et al., 2025) | | ✗ | | & | ✔ | ✗ | ✓ | 644 | 428.9 | 2,814 |
| **Live-Aid (Benchmark)** | | ✓ | | & | ✓ | ✓ | ✓ | **1,307** | **493.0** | **6,459** |

Table 7: **Comparison between Live-Aid and existing video benchmarks.** ***MU***, ***DM***, and ***TR*** denote Multimodal Understanding, Dialogue Modeling, and Temporal Reasoning capabilities, respectively. ***#Clips*** indicates the number of video clips. ***Len(s)*** is the average video duration.

| Capability | Samples |
|---|---|
| **Multimodal Understanding (MU)** | **2,099** |
| Video QA | 668 |
| - MCQ4 | 538 |
| - Open-ended | 130 |
| Multimodal Consistency | 686 |
| - True/False | 181 |
| - MCQ2 | 241 |
| - MCQ4 | 264 |
| Danmaku Intent Prediction | 745 |
| **Dialogue Modeling (DM)** | **1,585** |
| Contextual Danmaku Generation | 877 |
| Structured Dialogue Summarization | 708 |
| **Temporal Reasoning (TR)** | **2,155** |
| Danmaku Ordering | 620 |
| Clip Timeline Reconstruction | 866 |
| Video Evidence Grounding | 669 |
| - History | 164 |
| - Instant | 280 |
| - Delay | 225 |

Table 8: Statistics of tasks across different capabilities.

preserve key events while filtering noise.

**Temporal Reasoning (TR).** This capability evaluates the ability to navigate the non-linear time structure of live streams. We explicitly categorize temporal dynamics based on the interval between the textual input timestamp ($t_{txt}$) and the appearance of visual evidence ($t_{vis}$): *History* ($t_{vis} < t_{txt}$), *Instant* ($0 \leq t_{vis} - t_{txt} \leq 30s$), and *Delay* ($t_{vis} - t_{txt} > 30s$). We introduce three tasks: (1) *Danmaku Ordering* addresses asynchronous communication by requiring the model to restore the relative temporal order of shuffled danmaku based on visual triggers. (2) *Clip Timeline Reconstruction* operates at the global level, requiring the reconstruction of the logical narrative flow from scrambled video clips of a single session. (3) *Video Evidence Grounding* demands precise localization, predicting the start and end timestamps of a specific event described by a textual query.

## B.2 Hierarchical Multi-Agent Collaborative Framework

We elaborate on the **Multi-Agent Collaborative Generation Framework** employed to construct *Live-Aid*. To transform raw human annotations into a difficulty-stratified and rigorous benchmark, we design a pipeline comprising four sequential phases: Modality Enrichment, Component Generation, Task Synthesis, and Verification. Detailed visualizations of the specific prompts or workflows for representative agents are provided in Figs. 8–11.

**Phase 1: Modality Enrichment Agent Group.** This phase functions as the perception foundation, ensuring that textual annotations are rigorously grounded in multimodal signals. It integrates three specialized roles: first, the *Timeline Alignment Agent* employs deterministic rules to align the timestamp of each Danmaku interaction with the precise video frame index, establishing a unified temporal coordinate system. Second, the *Visual Context Enricher* utilizes Qwen3-VL-30B-A3B, a strong vision-language model, to generate dense, time-coded captions that capture scene transitions, host gestures, and on-screen text (OCR), thereby creating a rich textual context $C_{visual}$. Finally, the *Metadata & Dependency Analyzer* extracts global session metadata and classifies the "Modality Reliance" of each interaction to distinguish between Audio-Dependent and Multi-Modal Dependent samples.

**Phase 2: Component Generation Agent Group.** Acting as a pre-processing engine, this phase constructs the semantic building blocks and difficulty-controlling elements required for task synthesis. The *Summary Refiner & Intent Agent*, leveraging GPT-4.1, rewrites colloquial human summaries into structured abstracts and dynamically evolves a *Seed Local Intent Set* to cover real-world user

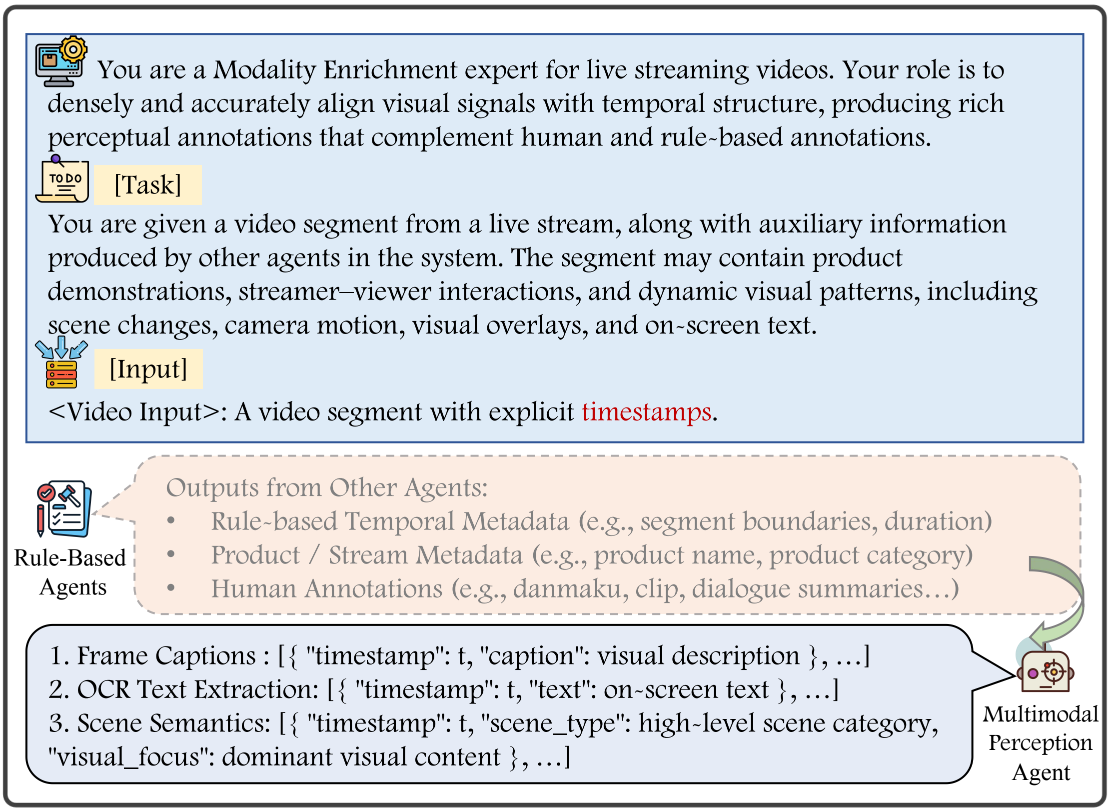



Figure 8: Prompt template and workflow for Modality Enrichment Agent Group.

behaviors. Crucially, to establish robust difficulty levels, the *Adversarial Distractor Agent* generates negative samples using a stratified strategy: it selects *Global Negatives* from entirely different live sessions and harder *Local Negatives* from the same video but distinct time windows, forcing the model to perform fine-grained temporal reasoning rather than relying on background scene bias.

**Phase 3: Task Synthesis Agent Group.** This phase acts as the central integration hub that synthesizes final executable tasks based on signals from Phase 1 and components from Phase 2. The agent receives a composite context including the Dialogue History $\mathcal{H}_{<t}$, current Danmaku, enriched visual captions $C_{visual}$, and generated distractors. Utilizing GPT-4.1, it integrates these inputs to formulate tasks across three core dimensions: *Multimodal Understanding (MU)*, *Dialogue Modeling (DM)*, and *Temporal Reasoning (TR)*. The output is formatted into standardized structures strictly adhering to task definitions, such as multiple-choice questions for Consistency, JSON objects for Summarization, and time intervals $[t_{start}, t_{end}]$ for Grounding.

**Phase 4: Verification Agent Group.** To minimize hallucinations and ensure objective solvability, we implement a rigorous two-step verification process. First, the *Consensus Committee*—comprising GPT-4.1, Qwen3-8B, and Qwen3-VL-8B—independently attempts to solve the generated tasks without ground truth; a sample is retained only if the committee reaches a consensus, proving the task is uniquely deducible. Second, a *Human-in-the-Loop Refinement* step serves as the final quality gate, where expert annotators eliminate subtle ambiguities and rectify domain-specific issues, such as cultural misinterpretations or precise frame-level boundary adjustments.

### B.3 Qualitative Examples

We present qualitative examples in Fig. 22–Fig. 30 to intuitively demonstrate the data format, input modalities, and annotation details of our benchmark, covering all tasks described in Appendix B.1. These selected cases highlight the intrinsic challenges of the dataset, particularly in scenarios requiring complex temporal reasoning and cross-modal grounding where visual cues are indispensable for correct deduction.

## C Additional Experimental Details

### C.1 Implementation Details.

**GPT-4o-mini, GPT-4o, and GPT-4.1:** Due to API limitations, we uniformly sampled 50

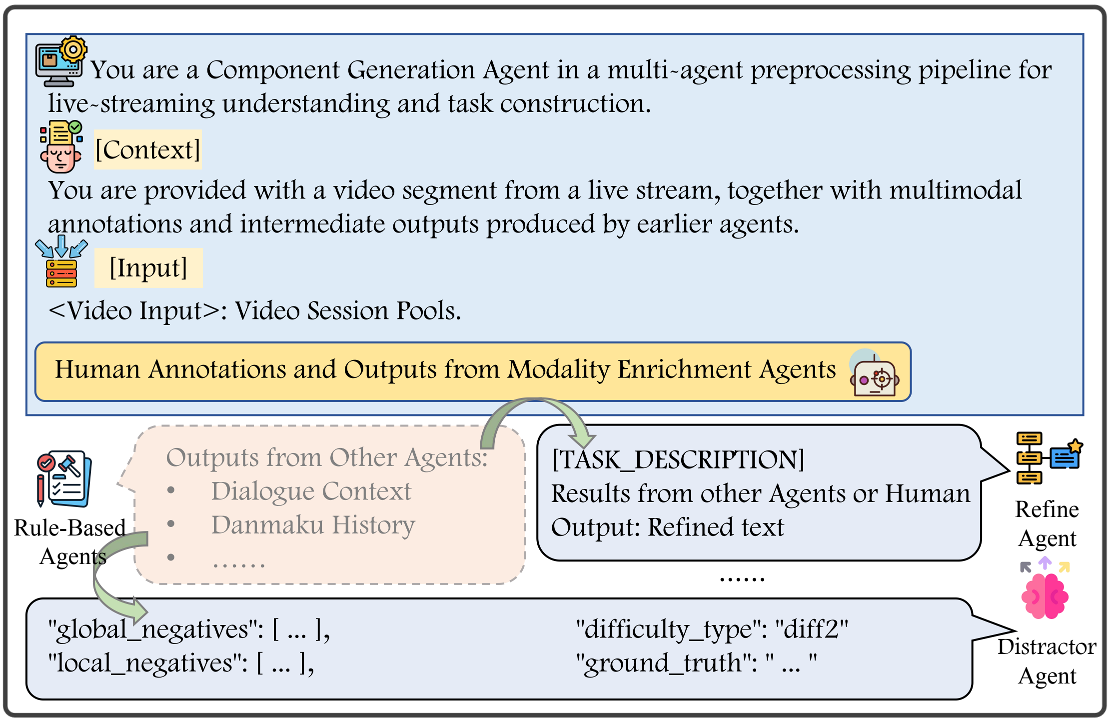



Figure 9: Prompt template and workflow for Component Generation Agent Group.

frames for videos longer than 50 seconds, while video clips shorter than 50 seconds were sampled at 1 fps. The model input follows the format: “<prompt_0><frames_0>...<frames_n-1> <prompt_n>”.

**Open-Source Models:** For vision models, the input follows the format “<frames> + <prompt>” for single-video tasks. For interleaved tasks involving multiple videos and text, the format “<prompt_0><frames_0>...<frames_n-1> <prompt_n>” is used. For omni models, we investigated the impact of audio by extracting audio tracks and incorporating them into the input. The format is “<frames> + <audio> + <prompt>” for single-video tasks, and “<prompt_0><frames_0><audio_0>... <frames_n-1><audio_n-1><prompt_n>” for interleaved tasks. Unless specified otherwise below, we set the default sampling rate to 1 fps and the maximum frame count to 32. Specific configurations for models with distinct sampling strategies are detailed as follows.

**Human Baseline:** We developed a dedicated annotation platform for human assessment, adopting a setup similar to that of MathVista (Lu et al., 2024). We randomly sampled 5% of the instances from the discriminative tasks and asked two human participants to answer the questions after watching the corresponding live-stream clips.

**Qwen3-VL:** Videos shorter than 768 seconds are sampled at 1 fps. For videos longer than 768 seconds, we extract 768 frames uniformly.

**InternVL3.5, VideoLLaMA2.1-AV, and OmniVinci:** We uniformly extract 32 frames regardless of video length. In cases where the frame rate is insufficient, VideoLLaMA2.1-AV employs padding, whereas InternVL3.5 and OmniVinci allow frame overlap/duplication to compensate.

**VideoLLaMA3:** Videos shorter than 128 seconds are sampled at 1 fps. For videos longer than 128 seconds, we extract 128 frames uniformly.

**MiniCPM-V-4.5:** Videos shorter than 96 seconds are sampled at 1 fps. For videos longer than 96 seconds, we extract 96 frames uniformly.

**LLaVA-Onevision(qwen2-ov):** We uniformly extract 32 frames for all videos. However, since frame duplication is disabled, the resulting frame count may be fewer than 32 if the video clip is too short.

### C.2 Evaluation Metrics.

We employ a diverse set of evaluation metrics tailored to the specific nature of each task, including standard accuracy, automatic metrics for discriminative tasks, advanced LLM-based semantic evaluations, and human evaluation.

**MultiChoice & Judgment Tasks.** For binary-choice, multiple-choice (four-option), and true–false questions, we adopt standard **Accuracy** as the evaluation metric to measure the percentage of correct predictions.

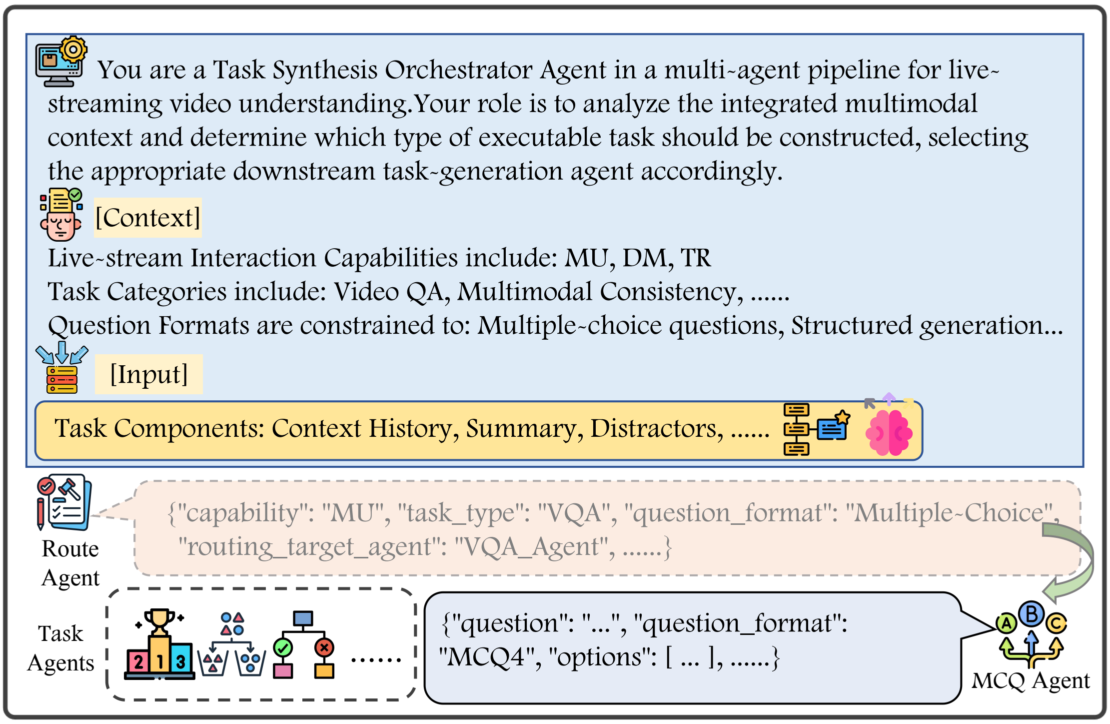



Figure 10: Prompt template and workflow for Task Synthesis Agent Group.

**Ranking Tasks.** For tasks requiring the restoration of a shuffled sequence, we utilize **Kendall's** $\tau$ and **Spearman's** $\rho$ as metrics. Kendall's $\tau$ measures the proportion of pairwise ordering agreements between the predicted and ground-truth rankings. Spearman's $\rho$ computes the correlation between the predicted and ground-truth ranks.

**Temporal Grounding Tasks.** For tasks requiring precise start and end timestamp localization, we employ temporal Intersection over Union (tIoU). This metric evaluates the quality of temporal alignment by computing the ratio of the intersection to the union between the predicted time segment and the ground-truth interval.

**Open-Ended Type Tasks (Automatic Metrics).** For generative tasks such as summarization and dialogue generation, we assess lexical and semantic quality using three standard metrics: BLEU-2, METEOR, and BERTScore F1. BLEU-2 measures bigram-level precision, while METEOR evaluates word-level alignment with stemming and synonym matching. BERTScore F1 computes semantic similarity using contextual embeddings to better capture meaning beyond surface-level overlap. For our Chinese dataset, BLEU-2 and METEOR utilize `jieba` for word segmentation, while BERTScore leverages `bert-base-chinese` for embedding-based semantic similarity.

**GPT-4.1-based Evaluation.** To complement n-gram metrics, which often fail to capture semantic fidelity and interaction quality, we introduce a GPT-4.1-based evaluation for open-ended tasks, scoring model outputs on a 0–10 scale with task-specific criteria. For *Video QA*, the evaluation focuses on correctness and visual grounding; for *Dialogue Summarization*, it emphasizes semantic completeness and factual consistency; for *Danmaku Generation*, it assesses intent alignment, contextual relevance, and stylistic naturalness in live-stream interactions. Hallucinations are strictly penalized across all tasks.

**Human Evaluation.** To complement automatic metrics, we conduct a human evaluation on a randomly sampled 5% subset of the data, following a protocol similar to the human baseline. We develop a scoring system to assess the quality of generated outputs on a 0–10 scale, where higher scores indicate better quality, ranging from perfect responses with no factual errors or hallucinations to severely flawed or irrelevant outputs. The evaluation prioritizes factual correctness over completeness and conciseness, with hallucinations strictly penalized.

### C.3 Evaluation Details.

For open-ended generative tasks, model outputs are evaluated directly without post-processing. For structured tasks, we apply task-specific regular expressions to extract the predicted answers from the raw model outputs. When multiple answer candidates are detected, the handling strategy depends

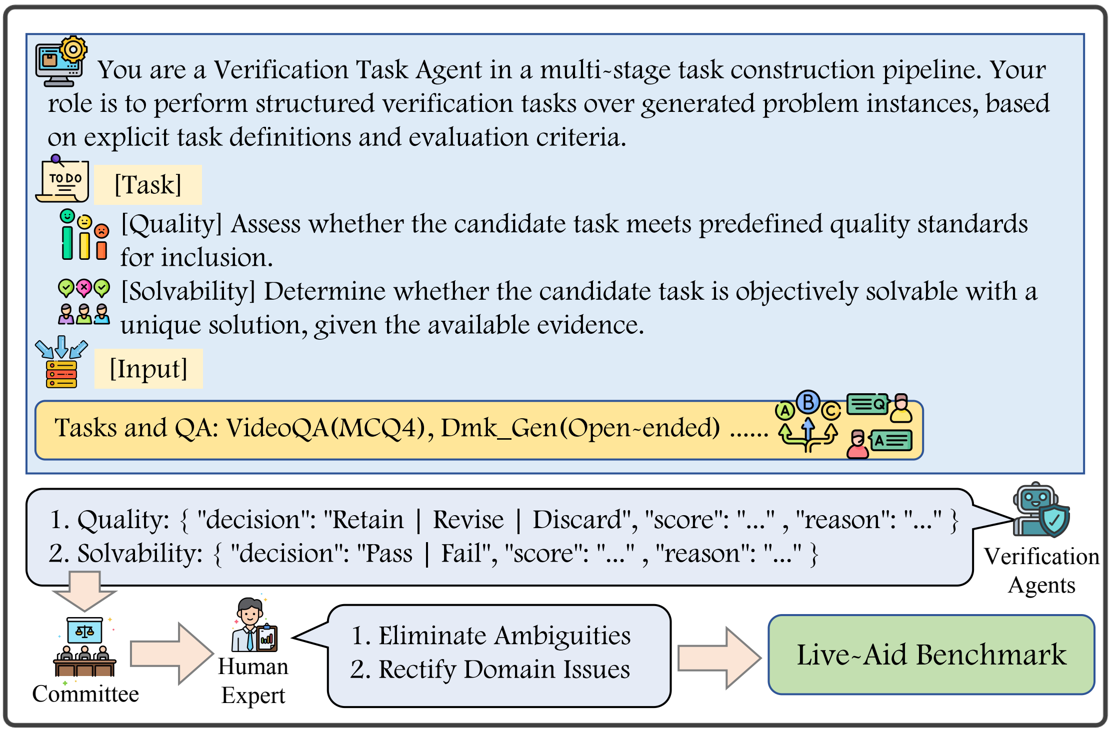



Figure 11: Prompt template and workflow for Verification Agent Group.

on the task definition. Specifically, for *Video Evidence Grounding*, where multiple temporal segments may validly constitute evidence, we merge all predicted time intervals (e.g., $[12.35, 16.76]$ and $[56.98, 72.31]$) into their mathematical union before computing the evaluation metrics. In contrast, for tasks requiring a unique solution, extracting multiple answers is treated as an incorrect prediction, as is the failure to extract any valid answer.

We also implement specific protocols for distinct task types. For *Danmaku Intent Prediction*, given the subjective nature of user intent, a prediction is considered correct if it matches any of the provided ground-truth labels. For *Ordering Tasks*, predictions with an incorrect number of elements undergo a penalty-based completion strategy, where missing elements are filled according to the reverse order of the ground-truth sequence. Additionally, we standardize outputs by normalizing 1-based indexing (e.g., 1, 3, 2) to 0-based indexing (e.g., 0, 2, 1) to ensure consistency. Finally, the detailed evaluation prompt templates used for the GPT-4.1-based assessment are illustrated in Fig. 12-Fig. 14.

# D More Results and Analysis

## D.1 Analysis of Distractor Difficulty

Tab. 9 reveals a clear performance hierarchy driven by the source of distractors. Models perform robustly in **Heterogeneous (HET)** and **Negative (NEG)** settings, where distractors originate from different videos and exhibit larger distributional gaps, making them easier to filter. In contrast, accuracy drops significantly in **Homogeneous (HOM)** and **Trap (TRP)** scenarios, where misleading options are drawn from the *same* video as the ground truth. This highlights the challenge of *intra-video discrimination*: same-source candidates share highly similar visual and semantic patterns, reducing the effectiveness of coarse-grained cues and introducing strong interference. As a result, models struggle to resolve fine-grained conflicts within a consistent visual context.

## D.2 Context Length Analysis

Fig. 15 demonstrates a general positive correlation between dialogue context length and model performance, as longer contexts provide a richer temporal horizon for modeling user intent evolution and recurring interaction patterns. However, performance fluctuations at extreme context lengths indicate challenges in utilizing long histories. As the number of turns grows, models must attend to more distant and potentially irrelevant information, which can dilute key signals and introduce noise; errors may accumulate over turns, leading to less stable predictions. This suggests a trade-off between leveraging richer contextual information and maintaining robust long-range reasoning.

**Prompt for Evaluation**

**System Prompt**
You are an expert in evaluating the quality of live-streaming product Q&A responses, tasked with assessing the quality of the [Model Prediction] based on the given [Question] and [Gold Answer].

**Evaluation Task**
Compare the model prediction with the gold answer, determine whether their semantics are consistent, and assign a score ranging from 0 to 10.

**Input**
- Question: {question}
- Gold Answer: {gold_answer}
- Model Prediction: {model_prediction}
- Item Title: {item_title} (for contextual reference **only**)

**Scoring Criteria (0–10)**
- 10 (Perfect): Semantically identical; all core information is correctly covered with no errors.
- 8–9 (Excellent): Core information is correct, with only very minor non-critical omissions or verbosity.
- 6–7 (Acceptable): Most key points are addressed, but some secondary information is missing or redundant.
- 3–5 (Poor): Core information is partially missing or contains some incorrect details.
- 0–2 (Incorrect): The answer is wrong, contradictory, irrelevant, or contains hallucinated content (e.g., fabricated attributes or parameters).

**Important Evaluation Rules**
1. Semantics First: **Do not** rely on surface-level string matching. Logical consistency and semantic equivalence are **prioritized**.
2. Hallucination Penalty: If the prediction introduces specific details or parameters not mentioned in the gold answer (e.g., fabricated numerical values), apply a severe **penalty**.
3. Colloquial Tolerance: More colloquial expressions in the prediction are acceptable.

**Output Format (Strict JSON)**
{
"score": <integer between 0 and 10>,
"reason": "A brief explanation in Chinese",
"label": "Correct" / "Partially Correct" / "Incorrect"
}

Figure 12: Evaluation prompt for the Video QA task.

| **Model** | **OA** | **MMC** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $MCQ_4$ | | | | $MCQ_2$ | | TF | | |
| | | HOM | HET | NEG | TRP | HOM | HET | POS | HOM | HET |
| Qwen3-VL$_{4B}$ | 53.9 | 39.0 | 55.0 | 51.6 | 20.1 | 60.7 | 79.1 | 35.1 | 57.5 | 89.3 |
| Qwen3-VL$_{8B}$ | 55.5 | 40.4 | 68.0 | 29.0 | 8.9 | 66.3 | 82.1 | 57.0 | 49.0 | 77.2 |
| InternVL3.5$_{4B}$ | 46.1 | 30.9 | 44.0 | 12.9 | 3.2 | 60.7 | 63.5 | 52.5 | 55.3 | 78.8 |
| InternVL3.5$_{8B}$ | 52.2 | 36.8 | 44.0 | 33.9 | 14.5 | 54.7 | 73.0 | 70.3 | 44.7 | 75.8 |
| Keye-VL-1.5 | 52.4 | 30.9 | 42.0 | 33.9 | 16.1 | 58.1 | 78.3 | 85.1 | 19.1 | 51.5 |
| VideoLLaMA3 | 49.0 | 25.0 | 28.0 | 67.7 | 74.2 | 47.0 | 52.2 | 25.7 | 70.2 | 84.8 |
| MiniCPM-V 4.5 | 54.5 | 35.3 | 68.0 | 25.8 | 4.8 | 65.0 | 83.5 | 67.3 | 40.4 | 63.6 |
| LLaVA-OneVision | 45.2 | 25.0 | 30.0 | 27.4 | 17.7 | 48.7 | 61.7 | 96.0 | 6.4 | 24.2 |
| LLaVA-OneVision-1.5 | 52.2 | 29.4 | 32.0 | 77.4 | 69.4 | 44.4 | 53.9 | 50.5 | 55.3 | 72.7 |
| VideoLLaMA2.1-AV | 58.2 | 20.6 | 38.0 | 100.0 | 100.0 | 48.7 | 59.1 | 81.2 | 19.1 | 24.2 |
| Qwen3-Omni-30B | 61.4 | 51.5 | 68.0 | 48.4 | 21.0 | 70.1 | 83.5 | 65.3 | 42.6 | 78.8 |
| OmniVinci | 57.3 | 33.8 | 40.0 | 66.1 | 61.3 | 54.7 | 74.8 | 24.8 | 95.7 | 100.0 |
| GPT-4o-mini | 48.7 | 32.4 | 52.0 | 12.9 | 6.5 | 54.7 | 69.6 | 85.1 | 21.3 | 57.6 |
| GPT-4o | 56.8 | 32.4 | 72.0 | 32.3 | 6.5 | 65.0 | 82.6 | 80.2 | 29.8 | 72.7 |
| GPT-4.1 | 60.5 | 41.2 | 78.0 | 25.8 | 4.8 | 76.1 | 85.2 | 91.1 | 23.4 | 60.6 |

Table 9: **Fine-grained performance on the MMC Task across difficulty settings.** We report accuracy (%) across varying levels of distractor similarity, including homogeneous (HOM/TRP) and heterogeneous (HET/NEG) sources.

### D.3 Analysis of Temporal Dependencies

Fig. 16 illustrates the performance of individual models across different temporal settings, revealing a clear stratification. While most models perform well in the *Instant* setting, their performance drops drastically in the *History* scenario. Notably, even strong baselines such as VideoLLaMA3 and LLaVA-OneVision-1.5 yield near-zero tIoU scores

**Prompt for Evaluation**

**System Prompt**

You are an expert in evaluating the quality of live-streaming dialogue summarization, tasked with assessing the model's performance by comparing the [Gold Summary] with the [Model Prediction].

**Task Background**

This task involves summarizing a live-streaming e-commerce dialogue.
- Gold Summary: A reference summary verified through human annotation or a high-quality curation process.
- Model Prediction: The summary generated by the model to be evaluated.
- Dialogue Context: Provided for reference **only**, and mainly used to detect hallucinations.

**Input**

- Item Title: {item_title}
- Gold Summary: {gold_summary}
- Model Prediction: {model_prediction}
- Dialogue Context: {dialogue_text}

**Scoring Criteria (0–10)**

Please assign a score based on the following dimensions:
1. Completeness (Core Information Coverage):
   Does the prediction include all **key points** from the gold summary (e.g., price, material, promotions, target audience)?
2. Faithfulness:
   Does the prediction introduce information that is not supported by the original dialogue (i.e., **hallucinations**)?
3. Conciseness and Fluency:
   Is the summary coherent, fluent, and free from unnecessary colloquial redundancy?

*Score Reference*:
- 10 (Perfect): Semantically identical to the gold summary; all core information is covered, with no hallucinations and fluent expression.
- 8–9 (Excellent): Covers all core information, with only minor wording differences or omission of extremely secondary modifiers that do not affect understanding.
- 6–7 (Acceptable): Captures the main selling points (e.g., product type and price) but omits secondary details (e.g., specific colors or size recommendations), or is slightly verbose.
- 3–5 (Poor): Misses core information (e.g., price or key attributes) or contains incorrect information.
- 0–2 (Incorrect): The summary is largely inconsistent with the facts, or contains severe hallucinations (fabricated attributes not present in either the gold summary or the dialogue).

**Output Format (Strict JSON)**

```
{
  "score": <integer between 0 and 10>,
  "reason": "A brief explanation in Chinese, specifying missing or incorrect information",
  "label": "Perfect" / "Acceptable" / "Poor"
}
```

Figure 13: Evaluation prompt for the Structured Dialogue Summarization task.

$(< 1\%)$. This sharp degradation effectively highlights the difficulty of *reverse temporal causality*, where models must retrieve relevant evidence that precedes the queried moment in time. Such a setting significantly weakens the effectiveness of forward temporal cues and requires integrating non-local information, often without reliable temporal anchoring, which remains highly challenging for current MLLMs and leads to unstable predictions.

## D.4 Analysis of Human Evaluation

Human evaluation primarily focuses on factual correctness, emphasizing the penalization of hallucinations, and assesses the overall quality of the generated outputs. Results from this evaluation show strong consistency with both the LLM-Score and automatic metrics, reinforcing the reliability of our framework. As illustrated in Tab. 10, top models like Qwen3-Omni and OmniVinci scored highly across all evaluation methods, demonstrating their robust multimodal interaction capabilities. Such alignment suggests the benchmark measures general multimodal interaction ability, rather than model-specific styles or biases. Consistency between human evaluation and other evaluation methods further highlights the robustness of our approach, ensuring an accurate reflection of model performance across various tasks.

## D.5 Impact of Audio Modality

The integration of audio generally yields consistent performance gains across most models,

| Model | Size | Frames | VQA | SDS | CDG |
|---|---|---|---|---|---|
| Open-source Video-LLMs | | | | | |
| Qwen3-VL-Thinking | 4B<br>8B | 1fps† | 4.95<br>4.40 | 5.95<br>5.40 | 4.65<br>5.65 |
| Qwen3-VL-Instruct | 4B<br>8B | 1fps† | 4.75<br>4.95 | 5.90<br>6.10 | 5.60<br>5.70 |
| InternVL3.5 | 4B<br>8B | 32 | 3.15<br>3.65 | 4.30<br>4.80 | 4.25<br>4.00 |
| Keye-VL-1.5 | 8B | 1fps† | 3.95 | 4.10 | 4.00 |
| VideoLLaMA3 | 7B | 1fps† | 3.90 | 4.40 | 4.10 |
| MiniCPM-V 4.5 | 8B | 1fps† | 4.00 | 5.10 | 4.65 |
| LLaVA-OneVision | 7B | 32 | 4.40 | 3.40 | 2.70 |
| LLaVA-OneVision-1.5 | 8B | 1fps† | 4.55 | 4.50 | 5.10 |
| Open-source Omni-LLMs | | | | | |
| VideoLLaMA2.1-AV<br>+audio | 7B | 32 | 4.25<br>3.20 | 3.35<br>2.65 | 4.05<br>2.55 |
| Qwen3-Omni-30B<br>+audio | A3B | 1fps† | 4.70<br>4.10 | 5.60<br>5.70 | 6.25<br>5.85 |
| OmniVinci<br>+audio | 9B | 32 | 4.80<br>6.10 | 4.95<br>6.60 | 5.60<br>6.80 |
| Commercial MLLMs | | | | | |
| GPT-4o-mini | ~8B | 1fps† | 4.90 | 5.75 | 5.30 |
| GPT-4o | ~200B | 1fps† | 3.75 | 5.55 | 5.80 |
| GPT-4.1 | - | 1fps† | 5.45 | 6.30 | 5.60 |

Table 10: **Human evaluation scores of model outputs on LiveAid across open-ended question types.** Scores (0–10) are based on overall generation quality and factual correctness.

confirming its role as a critical complementary source in information-dense live-streaming scenarios. Specifically, audio signals prove indispensable for resolving semantic ambiguities where visual cues alone are insufficient. However, our ablation results reveal that expanding modalities introduces significant challenges in **cross-modal alignment**. While top-performing models effectively synergize acoustic and visual signals, VideoLLaMA2.1-AV exhibits a *counter-intuitive performance drop* when audio is enabled compared to the vision-only baseline. This regression suggests that due to suboptimal multimodal integration, the audio stream acts as **interference noise** rather than informative context. Consequently, simply adding modalities does not guarantee improvement; without robust synchronization, it burdens the model's context capacity and disrupts the reasoning process.

## D.6 Impact of Sequence Length

Fig. 17 illustrates the performance trajectory across varying sequence lengths. Note that due to data sparsity for lengths $> 8$, these instances are aggregated into a single ultra-long category. At the shortest setting ($L = 3$), models exhibit noticeable variance, likely due to the limited availability of informative cues in short sequences. As the sequence length increases, performance follows an **inverted-U pattern**, peaking at lengths 6–7. This improvement suggests that mid-length sequences provide a "sweet spot": the enriched timeline offers sufficient contextual signals to reconstruct the narrative flow without overwhelming the model. However, beyond this threshold ($L > 7$), performance begins to decline. This drop indicates increasing difficulty in modeling long-range dependencies, where the accumulation of information makes it harder to accurately retain and order complex relationships.

## D.7 Impact of Evidence Granularity

While the *Clips* and *Full* settings in the Danmaku Ordering task share identical data samples, they differ fundamentally in evidence presentation, thereby evaluating distinct cognitive capabilities. The *Clips* setting provides explicit temporal boundaries, effectively reducing the task to local semantic alignment. Consequently, models with robust fine-grained grounding capabilities (e.g., MiniCPM-V 4.5) excel in this setting, mirroring their strong performance on the Multi-Modal Consistency task. Conversely, the *Full* setting relies on implicit temporal causality. Instead of matching pre-segmented clips, models must reconstruct the narrative flow directly from the raw video stream, a process that demands global event modeling and long-range reasoning. The performance gap observed between these settings highlights a critical insight: strong local alignment skills do not automatically translate to an understanding of global temporal structure.

## D.8 Case Study

Fig. 22-Fig. 30 provide a comprehensive qualitative analysis across diverse task types. The results reveal that while models excel at local semantic matching, the high-density, interleaved dynamics of live streaming pose a severe challenge. When reasoning is required over evolving dialogue contexts and fine-grained audio-visual dependencies, model robustness degrades significantly, leading to frequent hallucinations and temporal misalignment. This confirms the inherent difficulty of modeling real-time interactions.

## Prompt for Evaluation

**System Prompt**

You are an expert in evaluating the quality of live-streaming danmaku generation, tasked with assessing whether the model-generated danmaku is appropriate for the current live-streaming context and whether the expressed intent is accurate.

**Input**

- Item Title: {item_title}
- Dialogue History (History Context): {history_text}
- Gold Danmaku: "{gold_danmaku}" (Represents the real user's focus at the current moment)
- Model Prediction: "{model_output}"

**Guidelines for Using Dialogue History**

Please note that repeated questions are common in live-streaming scenarios due to new viewers continuously joining the stream.
- **Do NOT** penalize the model solely because a similar question appears in the earlier history.
- The primary role of the history is to help you identify the current **topic**
  (e.g., whether the discussion is about material, price, sizing, or promotions).

**Scoring Criteria (0–10)**

Please assign a score based on the following principles:
1. Intent Alignment (**Primary** Criterion):
  - Is the focus of the model-generated danmaku semantically aligned with the Gold Danmaku?
  - If the Gold asks "How much does it cost?" and the model asks "What's the price?", assign a full score.
  - If the Gold asks "Can we see the back?" while the model asks "How much is it?", this indicates a failure to follow the visual/contextual flow and should receive a low score.
2. Context Responsiveness:
  - Natural Repetition (Allowed): Repeating a question that was answered long ago is reasonable, as it simulates new viewers joining.
  - Invalid Repetition (Penalized): If the streamer explicitly emphasized information in the **immediately** preceding turn (Turn N−1), and the model redundantly asks the **same** question again in Turn N, this indicates poor contextual awareness and should be penalized.
3. Style:
  - The danmaku should be short, **colloquial**, and **viewer-like**. Overly formal or customer-service-style expressions should be penalized.

*Score Reference*:
- 10 (Perfect): Intent is fully aligned with the Gold danmaku, with natural and appropriate style.
- 8–9 (Excellent): Intent is reasonable and consistent with the current topic, though it may not exactly match the Gold's specific wording, and no obvious mistakes are made.
- 6–7 (Acceptable): A generic, non-specific danmaku (e.g., "Looks good", "Bought it"), or repetition of information discussed a long time ago.
- 3–5 (Poor): Intent is clearly misaligned (e.g., asking about shipping when others discuss sizing), or ignores information explicitly stated in the immediately previous turn.
- 0–2 (Very Poor): Hallucinated, nonsensical, irrelevant content, severe misunderstanding of context, or meaningless repetition.

**Output Format (Strict JSON)**

{
  "score": <integer between 0 and 10>,
  "reason": "A brief explanation in Chinese, comparing the intent of the Gold and the Model prediction and considering the dialogue history",
  "label": "Perfect" / "Acceptable" / "Context Error" / "Bad Style"
}

Figure 14: Evaluation prompt for the Contextual Danmaku Generation task.

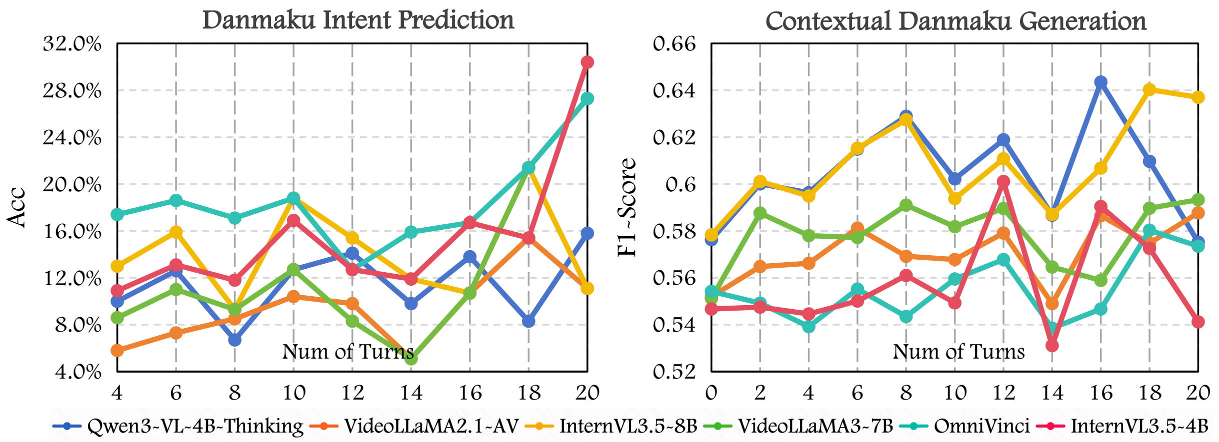



Figure 15: **Performance of MLLMs under different dialogue context lengths.** Results include Accuracy (%) for Danmaku Intent Prediction and BERTScore F1 ($\mathbf{F1_{BERT}}$) for Contextual Danmaku Generation.

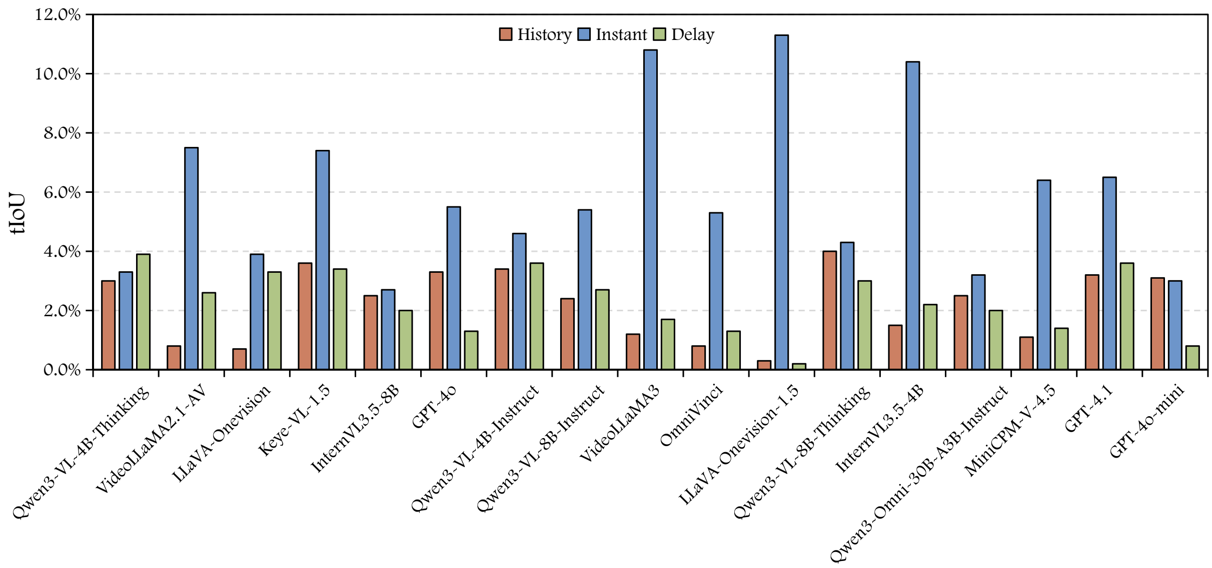



Figure 16: **Performance of MLLMs on the Video Evidence Grounding task across three temporal types.** Results are reported as IoU (%).

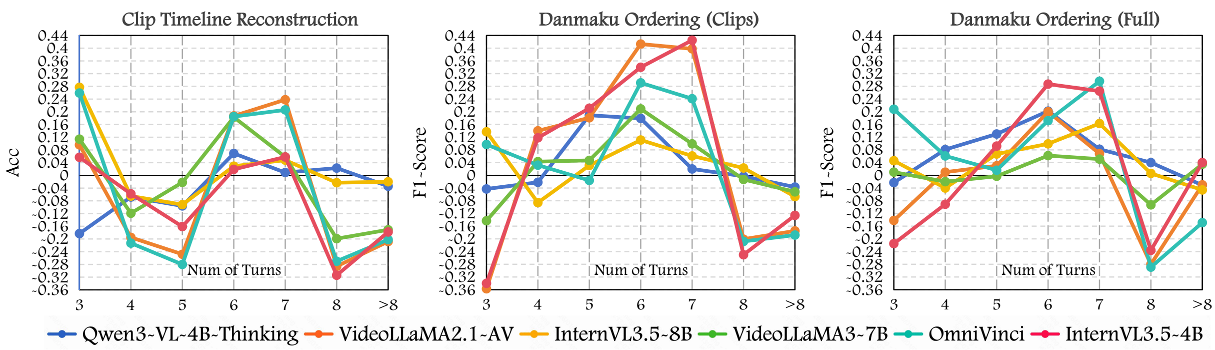



Figure 17: **Model performance on ordering tasks across different sequence lengths.** Results are reported as Kendall's $\tau$.

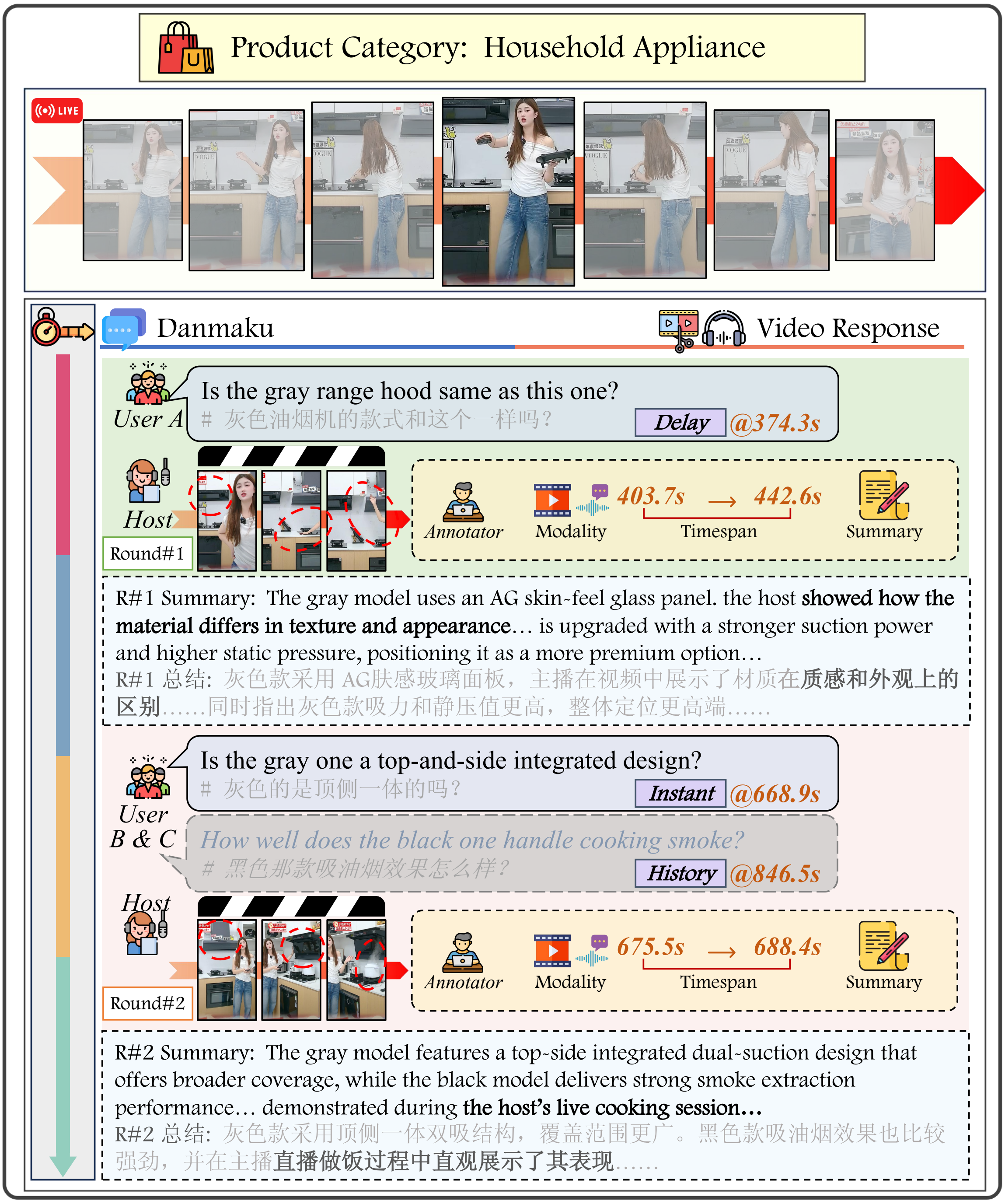



Figure 18: A representative sample from the Household Appliance category (range hood)


Figure 19: A representative sample from the Women's Shoes category


Figure 20: A representative sample from the Women's Apparel category


Figure 21: A representative sample from the Outdoor Gear category

Task: Video QA

Type: Open-ended

Danmaku

Question

User A: Do you have this in different colors?
# 有什么颜色的？ *Instant* *@138.1s*

What colors are available for this men's outdoor jacket?

R# Summary: The host showcased the color options for the outdoor jacket, which include purple, gray, and black.
R# 总结：主播展示了冲锋衣的颜色选项，包括紫色、灰色和黑色。 *134.4s ⟶149.2s*

**GPT-4.1:** This men's jacket is available in light purple and light gray. #这款男士冲锋衣有浅紫色和浅灰色两种颜色可选。

**GPT-4o-mini:** The available colors for this men's jacket include white and purple. #这款男士冲锋衣可选的颜色包括白色和紫色。

**GPT-4o:** Purple, white. #紫色、白色。

**InternVL3.5-4B:** Gray and light gray. #灰色和浅灰色。

**Keye-VL-1.5-8B:** Three colors. #三款。

**LLaVA-OneVision-1.5:** The men's jacket shown in the video is available in gray and black. #视频中展示的男士冲锋衣有灰色和黑色两种颜色可供选择。

**LLaVA-OneVision:** This men's jacket is available in white and black. #这款男士冲锋衣有白色和黑色两种颜色可选。

**MiniCPM-V-4.5:** Dark blue, purple, black. #深蓝色、紫色、黑色。

**OmniVinci (w/ audio):** Purple, gray, black. #紫色、灰色、黑色。

**OmniVinci (w/o audio):** This men's jacket is available in three colors: black, white, and gray. #这款男士冲锋衣有三种颜色可选: 黑色、白色和灰色。

**Qwen3-Omni (w/ audio):** According to the video, this jacket is available in three colors. #根据视频内容，这款冲锋衣有三个颜色可选。

**Qwen3-Omni (w/o audio):** According to the video, this jacket is available in light gray and white. #根据视频内容，这款冲锋衣有浅灰色和白色两种颜色可选。

**Qwen3-VL-4B-instruct:** Gray, black, white. #灰色、黑色、白色。

**Qwen3-VL-8B-Thinking:** The video does not mention color options for the men's jacket; only a light purple women's jacket is shown. #视频中未提及男士冲锋衣的颜色选项，仅展示了女士穿着的浅紫色冲锋衣。

**VideoLLaMA2.1-7B-AV (w/ audio):** According to the video, this men's jacket is available in three colors: gray, navy, and black. #根据视频，这款男士冲锋衣有三个颜色可选: 灰色、藏青和黑。

**VideoLLaMA2.1-7B-AV (w/o audio):** According to the video, this men's jacket is available in three colors: gray, navy, and black. #根据视频，这款男士冲锋衣有三个颜色可选: 灰色、藏青和黑。

**VideoLLaMA3-7B:** Black, gray, beige. #黑色、灰色、米色。

**Ground Truth:** Purple, gray, black. #紫色灰色 黑色

Figure 22: Qualitative case study of Video QA with an open-ended question type, including the model's output for analysis.

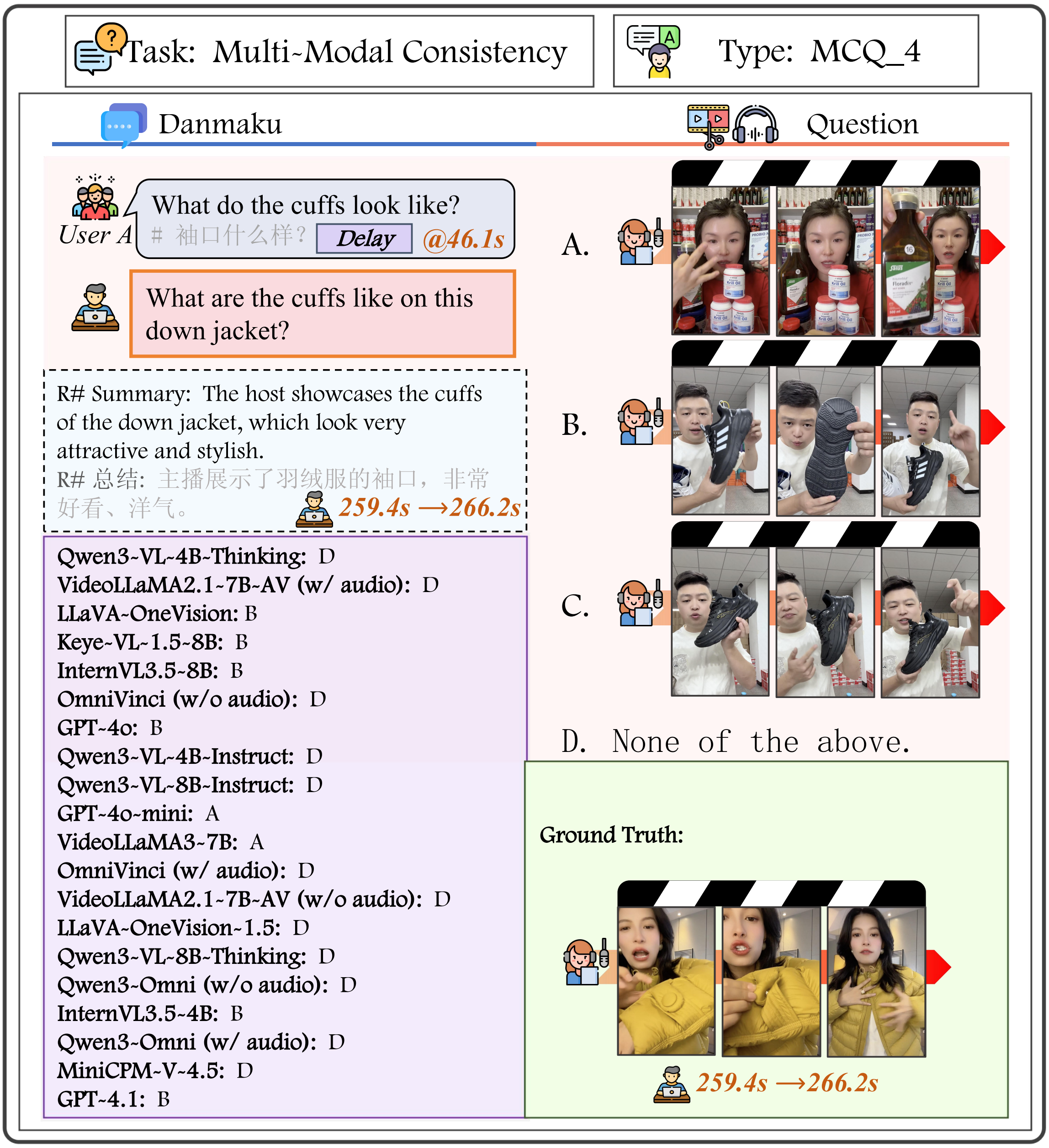



Figure 23: Qualitative case study of Multi-Modal Consistency with an multi-choice question type, including the model's output for analysis.

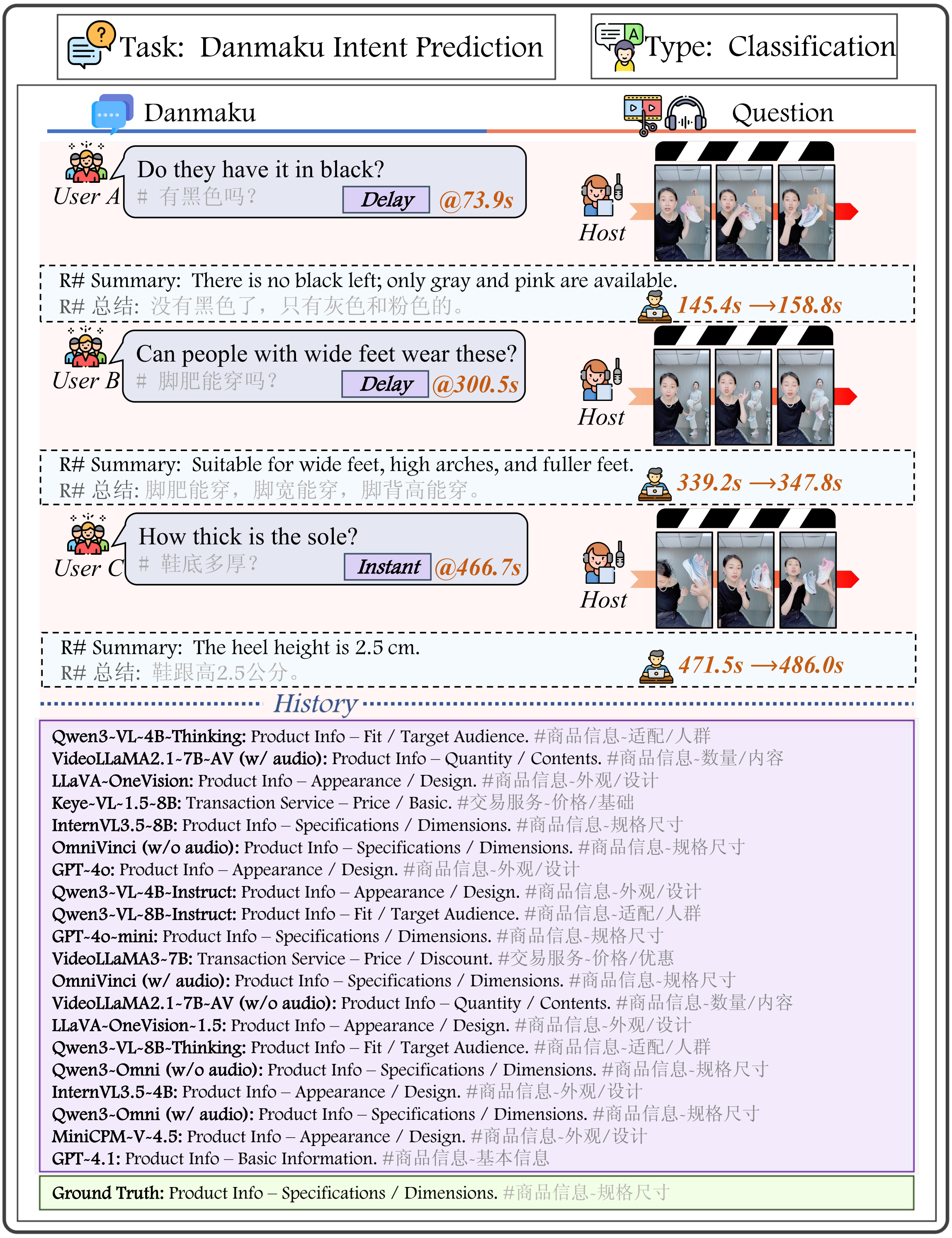



Figure 24: Qualitative case study of Danmaku Intent Prediction with classification question type, including the model's output for analysis.

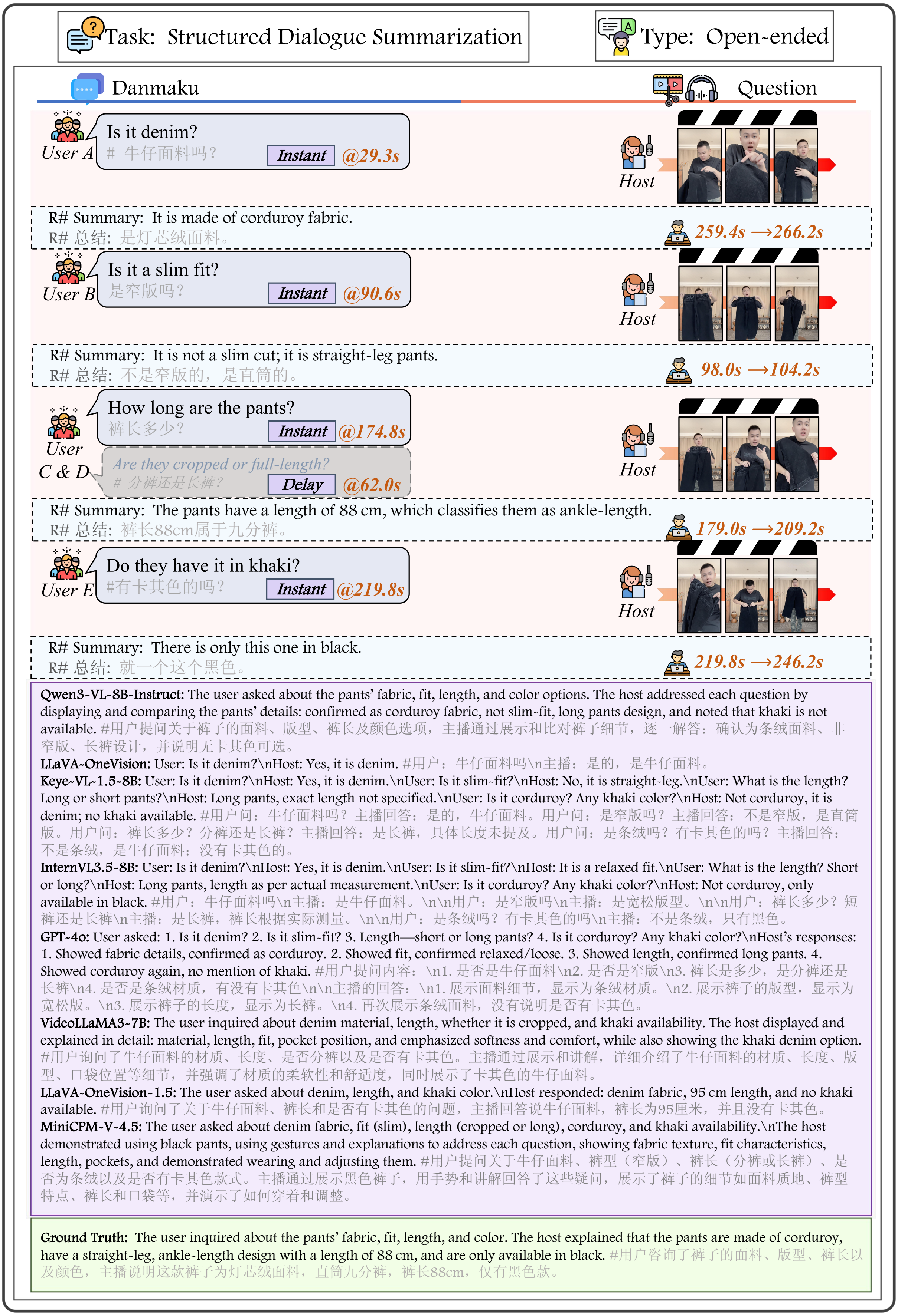



Figure 25: Qualitative case study of Structured Dialogue Summarization with open-ended question type, including the model's output for analysis.

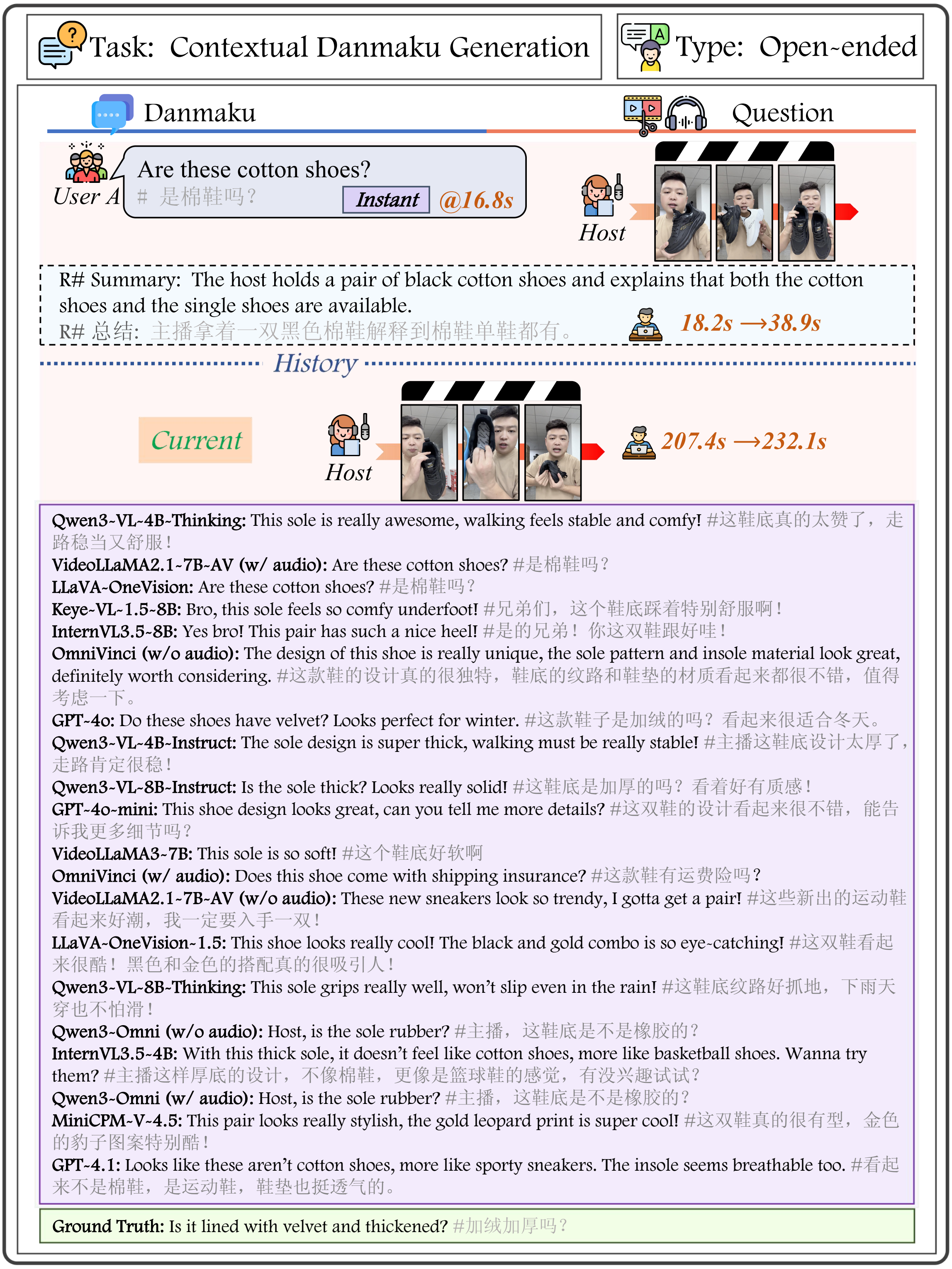



Figure 26: Qualitative case study of Contextual Danmaku Generation with open-ended question type, including the model's output for analysis.

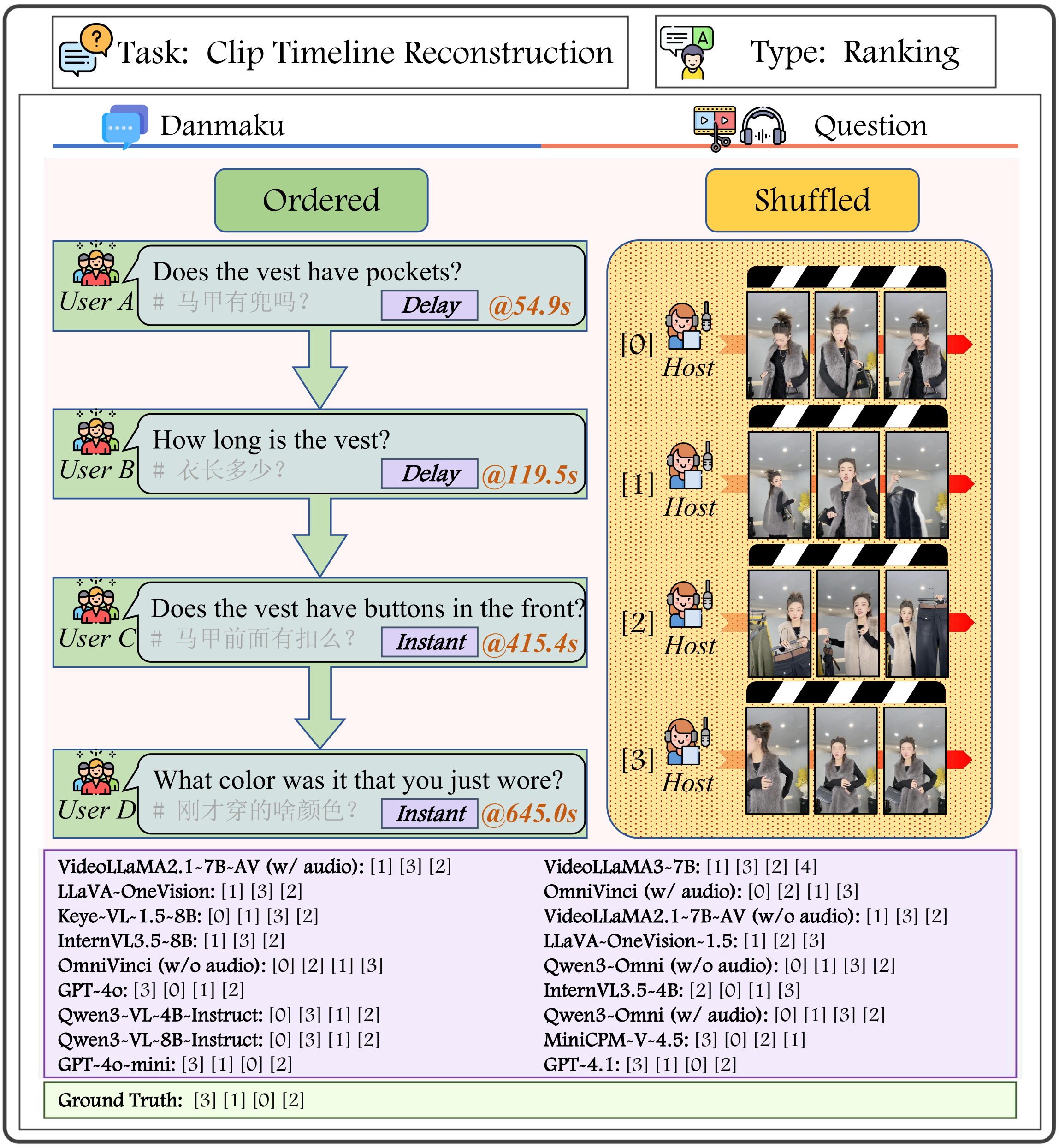



Figure 27: Qualitative case study of Timeline Reconstruction with ranking question type, including the model's output for analysis.

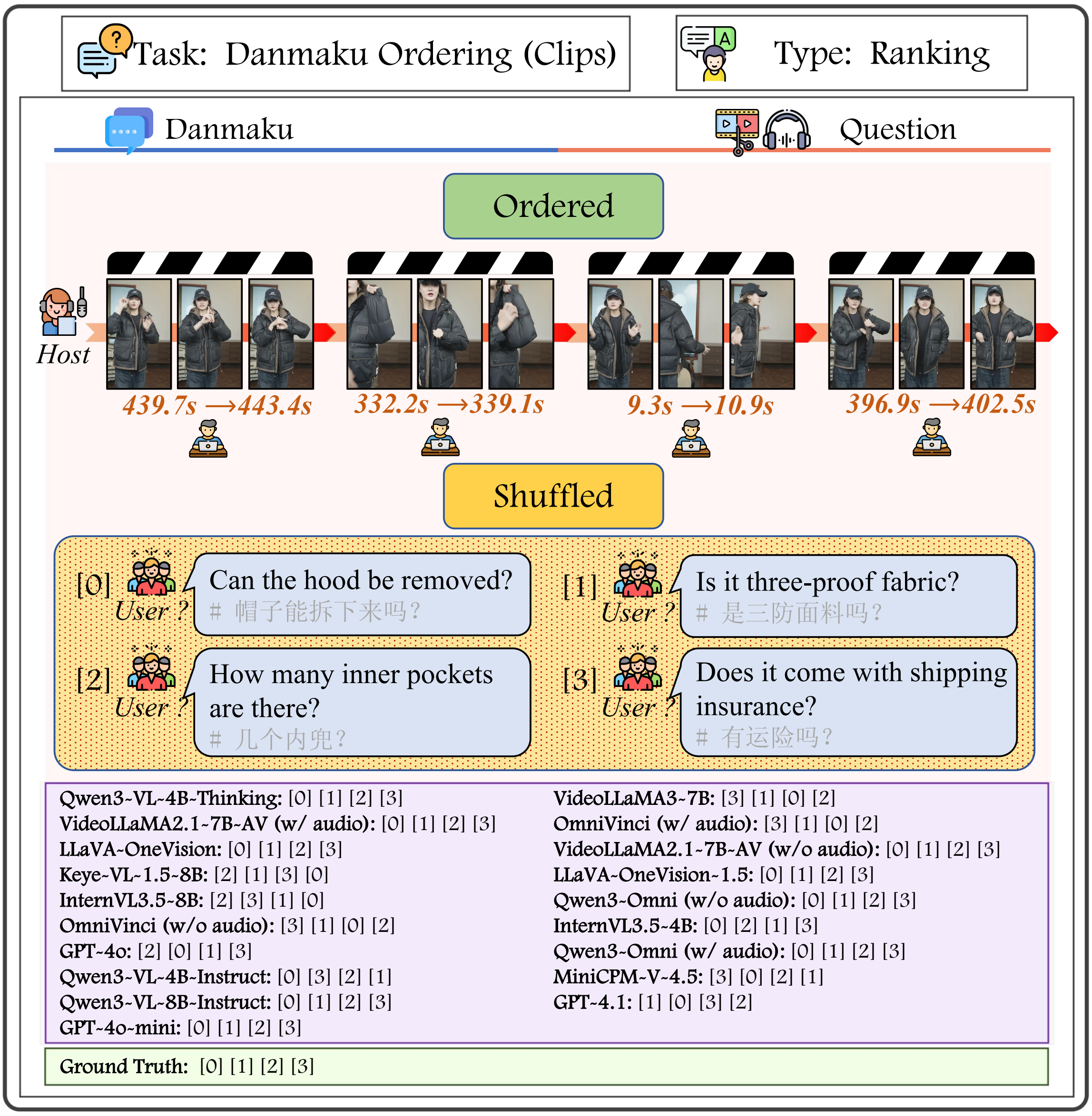



Figure 28: Qualitative case study of Danmaku Ordering (Clips) with ranking question type, including the model's output for analysis.

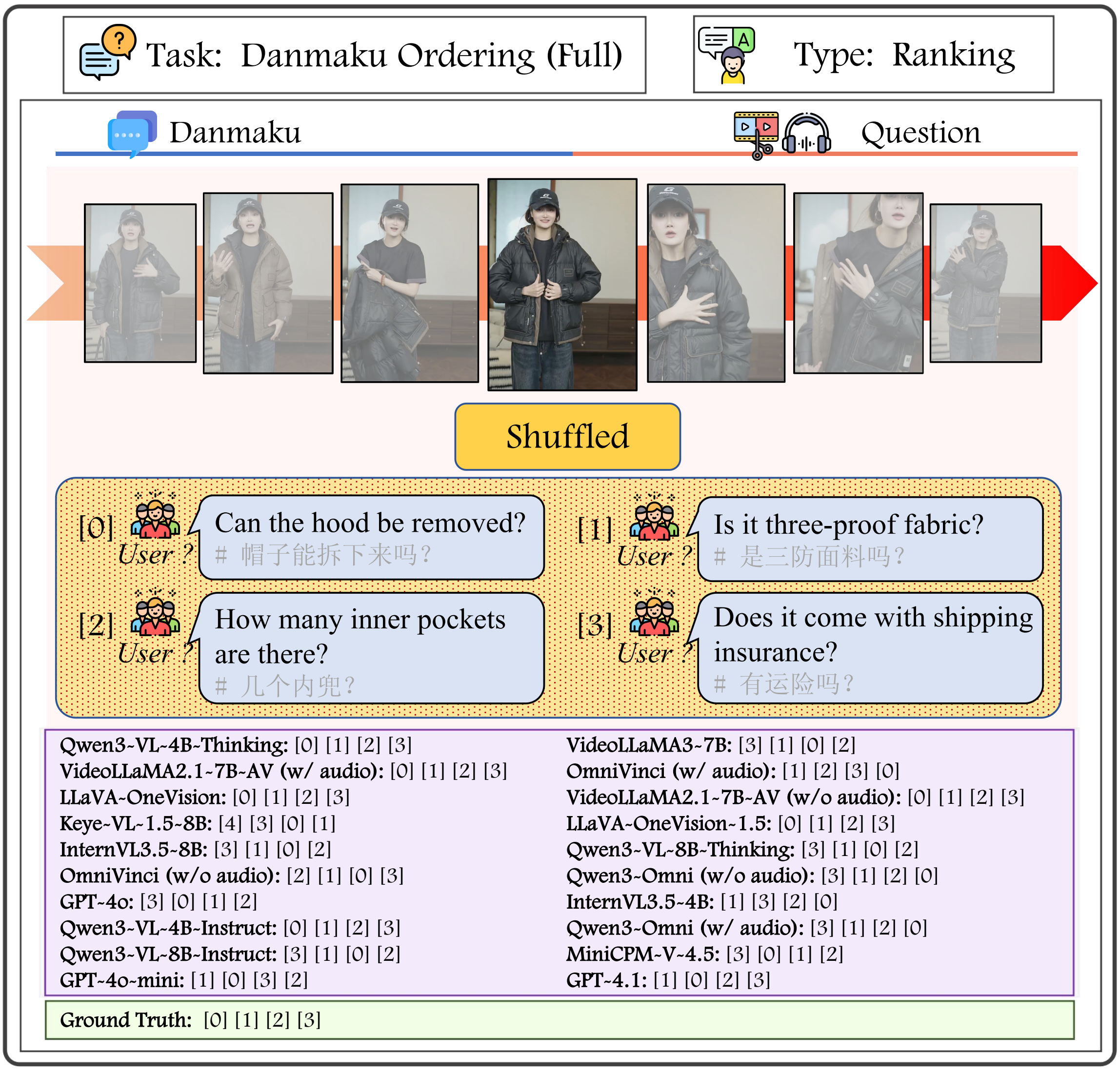



Figure 29: Qualitative case study of Danmaku Ordering (Full) with ranking question type, including the model's output for analysis.

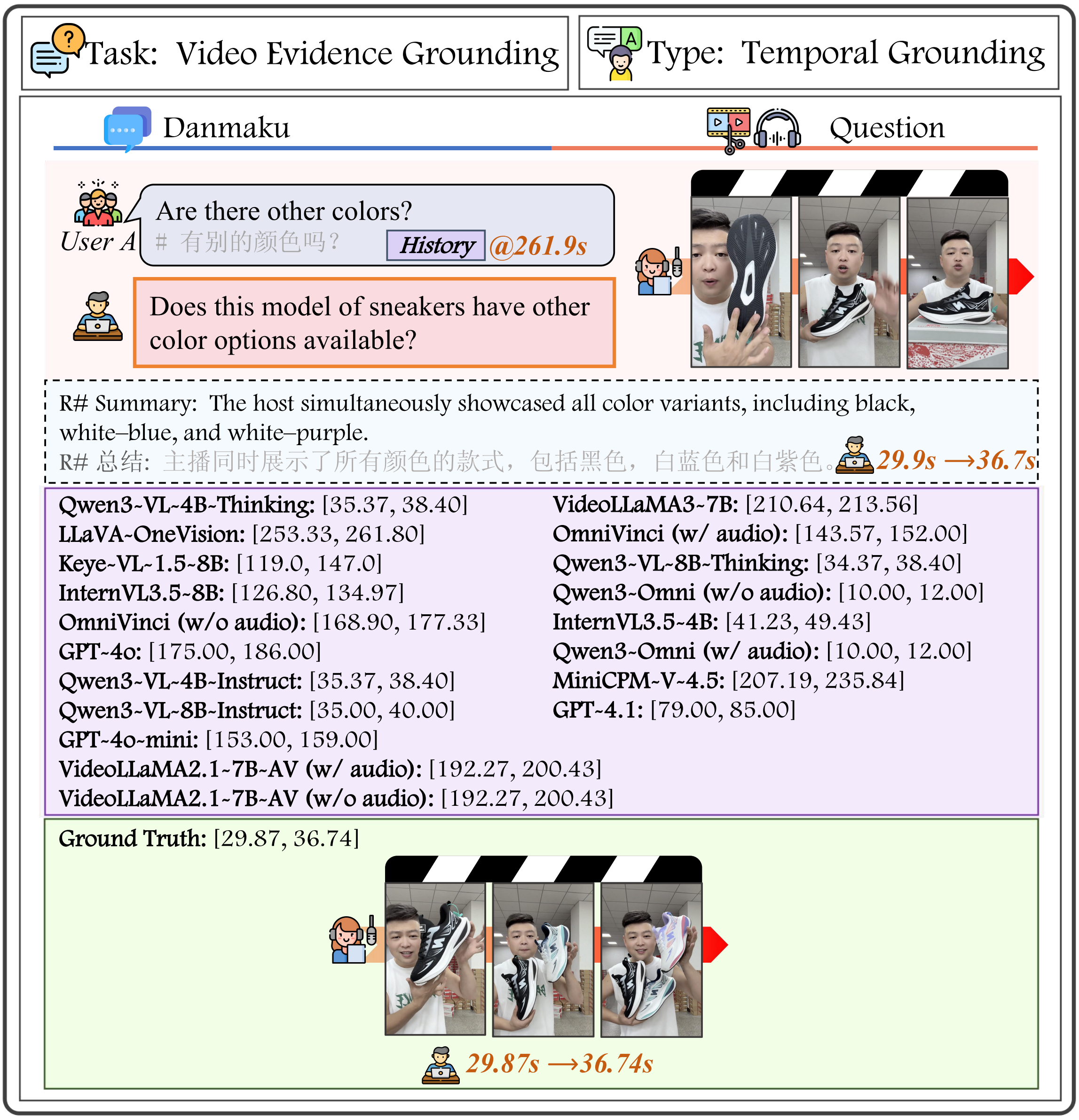



Figure 30: Qualitative case study of Video Evidence Grounding with temporal grounding question type, including the model's output for analysis.