GUESS WHAT YOU ASK: ANTICIPATING MOMENT-LEVEL REAL VIEWER INFORMATION NEEDS IN VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

A proactive video agent must decide what information to prepare before a viewer asks. We formulate a prerequisite task of proactive video agent as Moment-level Information Need Anticipation (MINA): given the video context observable up to a video timestamp, predict the questions real viewers are likely to ask. We construct MINA-Data from 12,510 timestamped questions naturally posted by viewers from 872 videos, and evaluate whether an answer to a predicted question would satisfy the observed need. Direct Forecasting is difficult (15.53% at best on Hit@5), even though most generated questions are semantically plausible. This reveals that the challenge is to select which of many plausible needs viewers actually express. We therefore propose MINA-GABT, which predicts a gap-formation state comprising an observed cue, the process by which it evokes an information gap, and the gap itself before verbalizing that gap as a question. A taxonomy induced from training-set viewer behavior guides state prediction. MINA-GABT reaches 21.20% Hit@5 and 27.12% Hit@10, with gains also on a held-out content category. Our analyses show that a video moment can support many plausible questions, but predicting which needs viewers will actually express remains difficult; event-specific gap formation and information beyond the video contribute to this uncertainty.

1 INTRODUCTION

Video question-answering (VideoQA) agents typically respond only after users explicitly ask a question, limiting their utility in natural viewing scenarios (Lei et al., 2018; Bodonhelyi et al., 2025; Cheema et al., 2026). A proactive agent should instead anticipate the possible information needs and prepare support before viewers express them (Rhodes & Maes, 2000; Liebling et al., 2012). This is particularly attractive for prerecorded content: with the video available in advance, such predictions and answers can be prepared offline, reducing online latency and interaction overhead. This motivates a new problem: given the video content observable to a viewer up to moment t, predict the information needs that real viewers are most likely to express at that moment. We refer to this task as **Moment-level Information Need Anticipation (MINA)**.

Table 1: Semantic plausibility does not imply behavioral anticipation. Despite producing predominantly plausible questions, both methods cover only a small fraction of the information needs actually expressed by viewers. Direct Forecasting uses Gemini-3.7-Flash, the gap-formation model in MINA-GABT.

	Semantic Plausibility	Behavioral Hit@5
Direct Forecasting	90.6%	12.35
MINA-GABT	94.4%	21.20

Studying this problem requires questions naturally produced by **real video viewers**, rather than synthetic or expert-written plausible questions (Yang et al., 2024a). We construct such a dataset from videos and their associated *danmaku* on Bilibili.[1] We identify danmaku that express information-seeking intent, yielding a dataset of 12.5K real viewer questions from 872 videos. To our knowledge,

[1]Danmaku are user comments synchronized to the video timeline, providing naturally aligned moment-level behavioral traces (Wu et al., 2018; Wang et al., 2020). Bilibili is a streaming platform characterized by Danmaku.

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

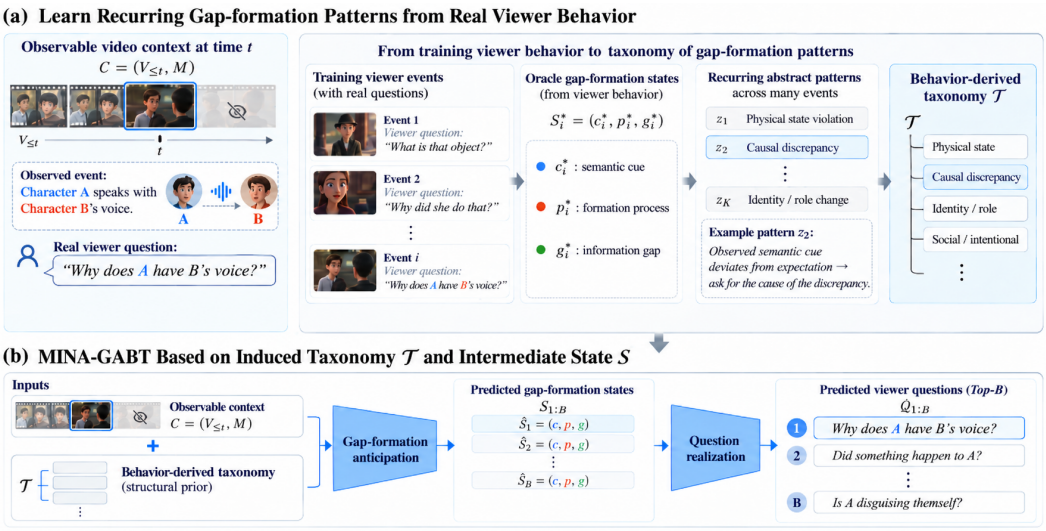

Figure 1: Overview of **MINA** (**M**oment-level **I**nformation **N**eed **A**nticipation) and **MINA-GABT**. Many natural-language questions are valid descriptions of plausible needs. We therefore move the anticipation target one level upstream, from question realization to the formation of the underlying information gap. (a) From real viewer questions in the training set, we annotate oracle gap-formation states S_i^* and induce a taxonomy $\mathcal{T}$ of recurring formation patterns. (b) Given the observable context C, MINA-GABT uses $\mathcal{T}$ as a structural prior to anticipate ranked gap-formation states $S = (c, p, g)$ and realize them as predicted questions. The repeated scene is illustrative; taxonomy induction uses training questions only and videos from the training split and test split do not overlap.

this is **the first moment-level video question dataset derived from naturally occurring viewer behavior**.

Evaluation shows that Direct Forecasting is difficult, but surprisingly, poor behavioral anticipation does not coincide with poor semantic plausibility. Our analysis finds that 90.6% of generated questions are semantically plausible, while behavioral Hit@5 remains very low. Moreover, a single moment can support 15.9 distinct plausible information needs on average. The challenge is therefore behavioral modeling: among many plausible questions, which information need will viewers actually express? This motivates us to model how a real viewer question emerges (Graesser & McMahen, 1993; Loewenstein, 1994). We model viewer-question emergence as a structured process, and introduce **gap-formation state** $S = (c, p, g)$ as a structured intermediate representation, where c denotes a semantic cue (e.g. anomaly, discrepancy, or salient contrast), p denotes the gap-formation process, and g denotes the resulting information gap from p. We reformulate forecasting as a two-stage process, where the model first predicts in the space of S and then verbalizes it as a natural-language question. We also introduce an agentic system that captures the recurring abstract patterns in the gap-formation process and induces a taxonomy $\mathcal{T}$ of such patterns. **The induced $\mathcal{T}$ serves as a structural, data-driven behavioral prior** learned from the observed viewer environment, and is used for predicting S. Combining these two designs, we introduce **MINA-GABT** (**G**ap-Formation **A**nticipation with a **B**ehavior-derived **T**axonomy). Experiments show that MINA-GABT consistently outperforms all baseline Direct Forecasting.

Our analyses further reveal that the remaining difficulty lies more in its context-specific instantiation than in identifying the abstract gap-formation pattern z. In addition, some predictive information is not contained in the observable video context, such as creator history and domain knowledge. The absence of such information is also a source of remaining difficulty. In summary, our contributions are fourfold:

- We formulate **MINA** (**M**oment-level **I**nformation **N**eed **A**nticipation): predicting the information needs real viewers will express from the video context observable at a given moment. We construct **MINA-Data** from 12,510 naturally occurring viewer questions across 872 videos and evaluate predictions by whether their answers can cover the viewer's need.

- We reveal a substantial gap between semantic plausibility and behavioral anticipation. Current VLMs predominantly generate questions that are plausible for the observed moment, yet perform poorly at predicting the particular needs real viewers express.
- We propose **MINA-GABT**, which anticipates a structured gap-formation state $S = (c, p, g)$ before realizing it as a question. A taxonomy induced from training-set viewer behavior supplies recurring formation patterns as a structural prior for predicting these states.
- MINA-GABT performs consistent improvements over Direct Forecasting. Analyses further locate the residual uncertainty primarily in event-specific gap formation rather than abstract-pattern identification, as well as information beyond observable video context.

2 Related Works

2.1 Proactive Assistance in Video Viewing

Interactive video assistants can provide contextual support during viewing, but often require viewers to formulate a question and initiate a request (Bodonhelyi et al., 2025; Cheema et al., 2026). This can interrupt viewing and viewers need to wait for responses. Such support can itself be distracting if delivered at unsuitable moments (Shin et al., 2018). These observations motivate preparing help in advance while retaining control over when to deliver it. JITIR and Anticipatory Search establish the idea of using context to surface useful information before explicit queries (Rhodes & Maes, 2000; Liebling et al., 2012). For streaming videos, assistants can continuously observe unfolding scenes and generate grounded responses (Chen et al., 2024; Zhang et al., 2025; Wang et al., 2026). Some benchmarks evaluate when to respond to a given question or instruction as a video unfolds (Wang et al., 2025; Chatterjee et al., 2026; Zhao et al., 2026). Epimopilot uses time-aligned comments to trigger contextual support for prerecorded videos (Qin et al., 2026). Prerecorded video offers a further opportunity, as assistance can be prepared offline to reduce online latency. The immediate question is therefore what information to prepare for a viewer at a given moment.

2.2 Video Question Answering and Generation

Video question answering (VideoQA) treats questions as inputs (Lei et al., 2018; Xiao et al., 2021). To automatically construct VideoQA datasets, video question generation (VideoQG) is designed to generate meaningful candidates grounded in video content (Yang et al., 2021), as well as entity-centric information-seeking questions (Phukan et al., 2024) and questions requiring reasoning across video segments (Phukan et al., 2025). These objectives establish ways to produce plausible questions, but do not by themselves determine which question a viewer will actually ask. MINA focuses on the latter problem. Naturally occurring viewer questions are also studied. YTCommentQA studies whether questions in instructional-video comments can be answered from the video (Yang et al., 2024a), while AQuA studies questions about software tutorials and how to answer them using visual anchors (Yang et al., 2024b). However, in these works, viewers' questions are not prediction targets but task inputs.

2.3 Behavioral Modeling of Viewer Information Needs

Among the questions a video moment can plausibly support, proactive agents must prioritize information needs that viewers in the target environment are likely to express. INQUISITIVE finds recurring pragmatic strategies in questions elicited during reading (Ko et al., 2020), suggesting that expressed viewer questions may also share reusable gap-formation patterns. For video, JIR-Arena evaluates time-dependent information needs (Yang et al., 2025). However, the proposed benchmark is based on expert annotation and LLMs, rather than naturally occurring questions. It is also limited to videos of lectures and conference talks. Naturally occurring platform interactions provide evidence of expressed behavior. Danmaku on Bilibili exhibit different behavior patterns from conventional comments (Wu et al., 2018). VideoIC uses moment-aligned danmaku to study video comment generation (Wang et al., 2020), while DanmakuTPPBench studies the timing of danmaku events (Jiang et al., 2026). These works show how audience behavior can be observed or modeled, but do not directly predict which information-seeking question a real viewer will express at a given video moment, which MINA focuses on.

3 MINA: TASK FORMULATION AND DATASET

3.1 TASK FORMULATION

We study **Moment-level Information Need Anticipation (MINA)**, where an agent predicts the information needs that real viewers are likely to express at a particular moment during video consumption. Formally, each event is represented as $e = (V, M, t, Q)$, where V is a video (typically prerecorded), M is the video metadata (e.g. the video title and description), t is the video timestamp at which a viewer expresses an information need, and Q is the naturally occurring viewer question. Q can be viewed as a sample from a viewer behavioral distribution:

$$Q \sim P_{\text{viewer}}\left(Q \mid C, \mathcal{E}\right), \tag{1}$$

where $\mathcal{E}$ denotes the behavioral environment in which viewing takes place, such as community culture and the creator ecosystem (Stewart & Mihalcea, 2022). $\mathcal{E}$ is not explicitly provided.

The agent observes $C_t = (V_{\leq t}, M)$, where $V_{\leq t}$ is the video prefix up to timestamp t. Although a prerecorded video can be processed offline, its prediction must respect the temporal video boundary. For brevity, we use C to denote C_t hereafter. Given C, the model produces a ranked list of B anticipated information needs in natural-language question form $\hat{\mathcal{Q}}^{(B)} = \left(\hat{Q}_1^{(B)}, \hat{Q}_2^{(B)}, \ldots, \hat{Q}_B^{(B)}\right)$.

3.2 MINA-DATA

Data Collection. We construct the MINA dataset from videos and their synchronized *danmaku* streams on platform Bilibili. Unlike questions written by annotators or generated by LLMs, danmaku are comments posted by viewers during natural video consumption and are aligned with the playback timeline, enabling collection of naturally emerging information needs and their corresponding video timestamps. For each video, we collect the video content, its title and description, and the associated timestamped danmaku stream. We retain a comment only if it expresses a genuine information-seeking intent and the intended information need can be reliably determined from the video context. Each retained event is represented as $e = (V, M, t, Q)$.

Canonical Information-Need Annotation. Naturally occurring danmaku often contain deixis, colloquial expressions, or references whose interpretation depends on the video context. To clarify the intended information need, we associate each question Q with a canonical annotation X^*. X^* is designed to be answer-equivalent to Q rather than lexical identical. Specifically, let $A(\cdot)$ denote the complete answer required to satisfy an information need. We require $A(X^*) = A(Q)$. X^* serves as an auxiliary semantic normalization that makes the information need explicit and self-contained for evaluation (Elgohary et al., 2019). Unless otherwise specified, all the evaluation and oracle analysis are conducted with access to the canonical X^* associated with the original question Q.

Video-Level Dataset Splits. We construct three splits: train, $\mathrm{test}_{\mathrm{ID}}$, and $\mathrm{test}_{\mathrm{OOD}}$. Splitting is performed at the video level, such that no video appears in more than one split. All learning is restricted to the training split, which can be further partitioned internally for training and validation. To separately evaluate in-distribution and topic-level generalization, we construct two test sets based on the platform-defined coarse content categories. Collected videos span 14 categories, including gaming, music, technology, sports, and animation. We hold out the entire category with ID 1005 (Animation/ACG) as $\mathrm{test}_{\mathrm{OOD}}$. Videos from the remaining 13 categories are divided between train and $\mathrm{test}_{\mathrm{ID}}$ while approximately preserving their within-category proportions.

Dataset Statistics. The resulting MINA-Data contains 12,510 naturally occurring information-seeking events from 872 videos, totaling approximately 136 hours of video content. Each video contributes 14.35 question events on average, with a median of 7, reflecting variation in the density of naturally occurring behavior. More statistics are in Table 2. We report human validation of MINA-Data in Appendix A.

3.3 EVALUATION PROTOCOL

MINA evaluates whether a predicted question would lead a proactive system to prepare information sufficient for the viewer's subsequently expressed need. We therefore evaluate predictions by the

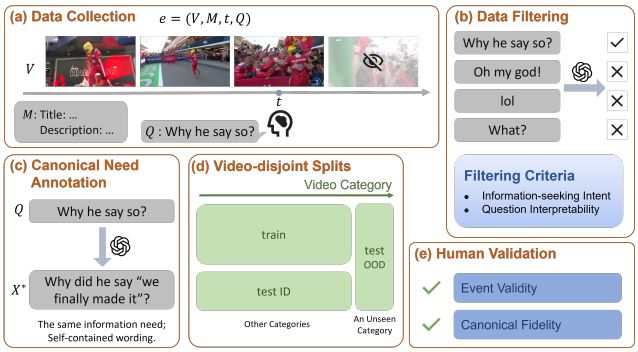

Figure 2: Overview of MINA-Data construction pipeline.

	Train	ID	OOD
# Videos	666	139	67
# Events	9,498	2,024	988
Pref. (Med)	186.98	259.84	212.67
Pref. (Avg)	491.05	678.23	360.47
# Events /V	14.26	14.56	14.75

Table 2: Statistics of the MINA dataset. "Pref." denotes the observable video duration (in seconds) before the target question timestamp. "# Events /V" denotes the averaged events count for each video.

information they request rather than by lexical or semantic similarity (Bulian et al., 2022). We define an asymmetric information-coverage relation as

$$Q_1 \succeq Q_2 \quad \Longleftrightarrow \quad A(Q_1) \text{ contains sufficient information to fully answer } Q_2. \tag{2}$$

The relation is intentionally asymmetric: $Q_1 \succeq Q_2$ does not in general imply $Q_2 \succeq Q_1$. For a predicted question $\hat{Q}$ and an observed viewer question Q, we define

$$h(\hat{Q}, Q) = \mathbb{I}\left[\hat{Q} \succeq Q\right]. \tag{3}$$

This criterion reflects the intended proactive use case of preparing information that can satisfy the information need before the viewer expresses it. This is implemented with an LLM-based evaluator. We report human validation of the evaluator in Appendix A.

4 MINA-GABT: Gap-Formation Anticipation with Behavior-Derived Taxonomy

Given the task defined in Section 3.1, we propose **MINA-GABT** (**G**ap-Formation **A**nticipation with a **B**ehavior-derived **T**axonomy) to anticipate a ranked set of information needs from the observable context C. MINA-GABT first induces a taxonomy of recurring gap-formation patterns from real viewer questions in the training set, and then uses the frozen taxonomy as a structural prior to anticipate gap-formation states for a new context. Each anticipated state is finally realized into one natural-language question, producing the ranked question predictions required by MINA.

4.1 Gap-Formation State

A natural-language question is the final expression of an information need, rather than the information need itself. Instead of taking the natural-language question itself as the primitive object of anticipation, we explicitly represent the state from which the question emerges. We model question emergence of real viewers as a structured process:

$$C \to c \xrightarrow{p} g \to Q, \tag{4}$$

where C is the video context, c is a semantic cue observed in the context, g is the information gap sought by the viewer, the gap-formation process p describes how the cue gives rise to the information gap, and Q is the natural-language question. This decomposition is motivated by cognitive accounts of discrepancies with prior knowledge, perceived information gaps, and the formulation of questions from knowledge goals (Otero & Graesser, 2001; Loewenstein, 1994; Ram, 1991). Appendix B discusses the connection and its limits. We use a simple example to illustrate it. In a cooking video (C), an unlabeled ingredient is added (c). The observation of an unidentified ingredient gives rise to a need for its identity (p), with the ingredient's unknown identity being the resulting information gap (g). This gap is finally verbalized as the question "What did he just add?" (Q). We combine the three components into a **gap-formation state** $S = (c, p, g)$. This shifts forecasting from question wording to the structured state underlying question formation. The final question is generated only after a gap-formation state has been predicted.

Table 3: Main results on test_{ID}. Methods marked with "*" use a two-stage pipeline with Gemini-3.7-Flash for need/state anticipation and GPT-5.6-Luna for question realization.

Method	Hit@5 ↑	MRR@5 ↑	Hit@10 ↑	MRR@10 ↑
Direct (Qwen3.8-Flash)	$8.15_{\pm.23}$	$4.72_{\pm.16}$	$13.09_{\pm.05}$	$5.82_{\pm.17}$
Direct (GLM-5.3-Flash)	$13.34_{\pm.22}$	$7.77_{\pm.18}$	$\mathbf{19.32}_{\pm.\mathbf{54}}$	$8.60_{\pm.36}$
Direct (Gemini-3.7-Flash)	$12.35_{\pm.40}$	$7.37_{\pm.28}$	$17.14_{\pm.22}$	$7.36_{\pm.12}$
Direct (Gemini-3.8-Flash)	$13.04_{\pm.28}$	$7.53_{\pm.19}$	$16.83_{\pm.67}$	$7.65_{\pm.35}$
Direct (Grok-4.7)	$14.36_{\pm.43}$	$8.37_{\pm.18}$	$\mathbf{19.32}_{\pm.\mathbf{10}}$	$8.78_{\pm.20}$
Direct (GPT-5.6-Luna)	$13.31_{\pm.34}$	$7.33_{\pm.09}$	$16.01_{\pm.56}$	$7.21_{\pm.30}$
Direct (GPT-5.6-Terra)	$13.14_{\pm.36}$	$7.73_{\pm.20}$	$18.17_{\pm.35}$	$7.78_{\pm.12}$
Direct (GPT-5.6-Sol)	$12.73_{\pm.35}$	$7.36_{\pm.18}$	$18.86_{\pm.25}$	$8.38_{\pm.14}$
Direct (GPT-6-Astra)	$\mathbf{15.53}_{\pm.\mathbf{34}}$	$\mathbf{8.66}_{\pm.\mathbf{28}}$	$19.25_{\pm.45}$	$\mathbf{8.93}_{\pm.\mathbf{18}}$
Free-form need only *	$16.67_{\pm.16}$	$9.82_{\pm.04}$	$22.48_{\pm.35}$	$10.10_{\pm.18}$
Intermediate S only *	$18.66_{\pm.23}$	$10.27_{\pm.08}$	$24.41_{\pm.53}$	$10.20_{\pm.30}$
Taxonomy $\mathcal{T}$ only (Gemini-3.7-Flash)	$15.17_{\pm.43}$	$9.02_{\pm.25}$	$23.09_{\pm.28}$	$10.05_{\pm.05}$
MINA-GABT *	$\mathbf{21.20}_{\pm.\mathbf{20}}$	$\mathbf{11.74}_{\pm.\mathbf{16}}$	$\mathbf{27.12}_{\pm.\mathbf{45}}$	$\mathbf{11.88}_{\pm.\mathbf{25}}$

4.2 Behavior-derived Gap-formation Taxonomy

The gap-formation state S provides an intermediate representation aligned with how viewer questions arise, while its gap-formation process p remains highly specific to the video context. However, many real viewer events share the same abstract structure underlying their formation process. For example, character A unexpectedly speaks with character B's voice, opening a need to explain the voice mismatch. This gap-formation process can be de-instantiated into a more general pattern: an observed state deviates from expected baseline, leading to an information need about the cause of the discrepancy. Such a pattern can describe variant events, such as a person unexpectedly behaving against their normal character. We denote such an abstract and reusable gap-formation pattern by z.

By extracting recurring patterns from the training data, we can construct a taxonomy $\mathcal{T} = \{z_1, \ldots, z_L\}$, which serves as a structural prior for anticipating context-specific gap-formation states. Patterns in $\mathcal{T}$ are non-exclusive, as a concrete formation process p may correspond to multiple patterns. The taxonomy is also intentionally non-exhaustive, to avoid overfitting to long-tail cases and degradation of the overall quality.

To induce such a taxonomy, we first annotate each training event with its oracle gap-formation state $S_i^* = (c_i^*, p_i^*, g_i^*)$ from the viewer question Q_i and its observable context C_i. We then introduce an induction harness to learn $\mathcal{T}$ from these annotated gap-formation states S_i^*, especially their formation process p_i^*. The agent can analyze the training evidence, maintain the evolving taxonomy, and evaluate it within the agent loop, but only through actions exposed by the harness. Some actions internally invoke LLMs, but they operate in separate contexts and serve only as stateless tools. The actions exposed by the harness are detailed in Appendix D. Through this harness, the agent model can efficiently maintain a data-consumption loop and revise the taxonomy as supported by the accumulated evidence. The validation set is obtained through a secondary split of the training split at the video level and shares no videos with either the data used for taxonomy induction or the benchmark test sets. The resulting taxonomy is committed only after its current version has been audited and validated. This procedure yields a compact set of recurring gap-formation patterns derived from real viewer behavior. The final taxonomy has $L = 37$ patterns. Refer to Appendix E for more details.

4.3 Taxonomy-Guided Gap-Formation Anticipation

Given a new observable context C and the frozen behavior-derived taxonomy $\mathcal{T}$, MINA-GABT first anticipates a ranked set of B gap-formation states:

$$\mathcal{F}_{\text{gap}}(C, \mathcal{T}; B) = \hat{\mathcal{S}}^{(B)} = \left(\hat{S}_1, \ldots, \hat{S}_B\right), \tag{5}$$

where each $\hat{S}_k = (\hat{c}_k, \hat{p}_k, \hat{g}_k)$ is a context-specific instantiation of a potential viewer information need. Each anticipated state is then independently realized into a natural-language question, and

Table 4: Generalization and model sensitivity.

(a) OOD Generalization

Method	Hit@5	MRR@5
Direct (Gemini-3.7-Flash)	$18.88_{\pm.31}$	$11.73_{\pm.10}$
Direct (GPT-5.6-Sol)	$18.88_{\pm.25}$	$11.54_{\pm.34}$
Direct (GPT-5.6-Luna)	$18.03_{\pm.27}$	$10.72_{\pm.21}$
Direct (GLM-5.3-Flash)	$17.19_{\pm.47}$	$10.22_{\pm.32}$
MINA-GABT	$25.80_{\pm.36}$	$15.75_{\pm.25}$
Intermediate S only	$23.74_{\pm.75}$	$14.30_{\pm.64}$
Taxonomy $\mathcal{T}$ only	$20.47_{\pm.63}$	$12.44_{\pm.41}$

(b) Model Sensitivity

Gap Model	Realization	Hit@5
Gemini-3.7-Flash	GPT-5.6-Luna	$21.20_{\pm.20}$
GPT-5.6-Sol	GPT-5.6-Luna	$21.38_{\pm.58}$
GPT-5.6-Luna	GPT-5.6-Luna	$18.86_{\pm.54}$
Gemini-3.7-Flash	GLM-5.3-Flash	$20.32_{\pm.53}$
Gemini-3.7-Flash	Gemini-3.7-Flash	$19.22_{\pm.20}$
Gemini-3.7-Flash	GPT-5.6-Sol	$20.32_{\pm.21}$

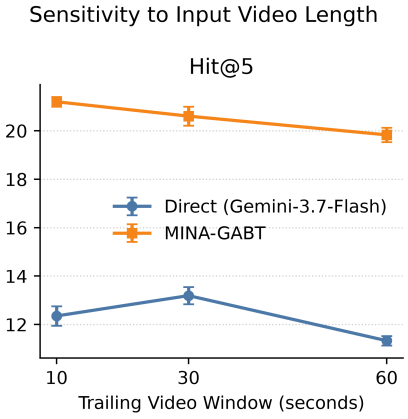

Figure 3: Sensitivity to the length of observable video history.

applying this function to all anticipated states yields the prediction:

$$\mathcal{F}_{\text{realize}}(C, \hat{S}_k) = \hat{Q}_k, \quad \hat{\mathcal{Q}}^{(B)} = \left(\hat{Q}_1, \ldots, \hat{Q}_B\right). \tag{6}$$

5 Experiments

We follow the evaluation protocol in Section 3.3. Refer to Appendix C for more experimental setups.

5.1 Main Results

Overall Performance. Even at their best, the evaluated VLMs achieve only 19.32% Hit@10. In contrast, MINA-GABT improves both the coverage and ranking of anticipated viewer needs on both prediction budgets. Relative to the strongest Direct Forecasting baseline, MINA-GABT improves Hit@5, MRR@5, Hit@10, and MRR@10 by 36.5%, 35.6%, 40.4%, and 33.0%, respectively.

Both separating need anticipation from question realization and explicitly modeling gap formation contribute to the improvement. To distinguish the proposed structured representation S from the more general benefit of two-stage need modeling, we introduce a free-form two-stage baseline, which predicts candidate information needs as free-form descriptions and then realizes them as questions. This free-form formulation improves Hit@5 from 12.35% to 16.67% and Hit@10 from 17.14% to 22.48%, supporting the benefit of explicitly modeling a latent information need before question realization. Explicitly structuring the intermediate representation as $S = (c, p, g)$ further improves performance to 18.66% Hit@5 and 24.41% Hit@10, showing an additional benefit of structured gap-formation modeling over a free-form need representation. An oracle analysis further shows that, given S^*, GPT-5.6-Luna can recover 83.53% of real viewer events with $B = 1$, indicating that the structured state is sufficiently expressive to preserve the target need. The behavior-derived taxonomy provides another complementary source of improvement: using $\mathcal{T}$ without S improves over Direct Forecasting, while combining $\mathcal{T}$ with structured gap formation yields the full MINA-GABT performance of 21.20% Hit@5 and 27.12% Hit@10. A post-hoc coverage analysis finds that the induced $\mathcal{T}$ covers 92.24% of unseen test events, supporting its role as a reusable structural prior.

These gains nevertheless leave absolute anticipation accuracy low. As analyzed in Section 6, this is not because the models fail to generate semantically plausible questions. Rather, **each moment admits many plausible needs, making the behavioral choice among them the central challenge**.

5.2 Generalization and Robustness

Category-Held-Out Generalization. We evaluate on test_{OOD}, whose videos all belong to a held-out category absent from training. We report results with $B = 5$ in Table 4. MINA-GABT retains an

advantage over Direct Forecasting, and both designs show gains over Direct Forecasting. The result indicates that the recurring gap-formation structure induced from training behavior is not restricted to the categories observed during taxonomy construction.

Sensitivity to the Foundation Model. We test whether the benefit of MINA-GABT depends on particular models. We replace the gap-formation prediction and realization model and report results in Table 4. Results with different realization models are stable, while replacing the gap model with GPT-5.6-Luna brings minor degradation. The reason might be that utilizing Taxonomy $\mathcal{T}$ brings extra difficulty for GPT-5.6-Luna, as its performance is similar to other models in the main results.

Sensitivity to the Observable Video History. MINA is defined over the causal video prefix $V_{\leq t}$, but practical systems may choose to expose only a recent temporal window for efficiency. We vary the length of trailing window while keeping the prediction moment and metadata fixed. For a trailing window of length ℓ, we define $C_t^{(\ell)} = \big(V_{[\max(0,t-\ell),\,t]}, M\big)$. Figure 3 shows that a longer trailing window typically brings degradation, except when comparing 30s and 10s windows for Direct Forecasting. A possible explanation can be that recent local context is enough for MINA, and a longer window might bring distractions.

6 Understanding The Behavioral Anticipation Gap

Main results establish the behavioral anticipation gap and show that gap formation and induced taxonomy partially closes it. We now ask where the remaining uncertainty comes from.

Semantic Plausibility vs. Behavioral Anticipation. A key source of difficulty in MINA is that generating a plausible question is different from anticipating the information needs that real viewers will actually express. We design a post-hoc analysis to validate the distinction and to demonstrate how it makes MINA challenging. We use a VLM-based auditor to test whether a predicted question from Direct Forecasting or MINA-GABT is plausible given the video context C. For each event, we also randomly sample predictions from other videos as negative controls to verify that the auditor is discriminative. Both methods predominantly generate plausible questions: 90.6% of Direct Forecasting and 94.4% of MINA-GABT predictions are judged plausible, while plausibility drops to 6.0% and 1.9%, respectively, on negative controls. Moreover, after de-duplication of 20 predicted questions (10 from Direct Forecasting, 10 from MINA-GABT), each event still contains 15.9 distinct plausible information needs on average (median 16), exceeding both evaluation budgets of 5 and 10. As de-duplication can be performed on a larger prediction pool, this figure only proves a lower bound of plausible questions. It suggests the context can support considerable plausible needs, and **the bottleneck of MINA is not the ability to yield plausible questions but to identify questions that viewers actually ask**.

Is the Structural Prior $\mathcal{T}$ Enough? The learned taxonomy $\mathcal{T}$ provides a reusable structural prior, but does not encode how viewer behavior is distributed over the taxonomy itself. We therefore augment it with either a population prior $p(z)$ or a context-conditioned prior $p(z \mid C)$, both learned from the training set. As shown in Table 5, surprisingly, even substantially improving the recovery of compatible abstract patterns z does not improve behavioral anticipation. This separates two sources of uncertainty: which general pattern applies and **how that pattern is instantiated into the concrete gap at this event**. The latter is substantially more limiting.

What Information Matters Beyond the Abstract Pattern? The plausibility analysis shows that the observable context admits many reasonable needs. We next ask what information can resolve this ambiguity toward the need actually expressed by the viewer. We perform a component-wise oracle analysis to answer this question. Specifically, we annotate $S^* = (c^*, p^*, g^*)$ and z^* for the test set, and evaluate performance when selected components are provided. Table 5 shows that providing z^* raises Hit@5 by 15.19 pp, confirming that the correct formation pattern is informative. Yet Hit@5 remains only 36.39%, showing that identifying the abstract pattern alone is far from sufficient. Furthermore, $c^* + p^*$ improves by 21.25 pp over $c^* + z^*$, reaching 77.65%. This means **the central residual challenge is not merely selecting an abstract behavior pattern, but instantiating that pattern into the particular information gap induced by the current event**. We

Table 5: Diagnosing the sources of residual uncertainty. (a) Additional priors over abstract patterns. (b) Component-wise oracle interventions on $\mathrm{test}_{\mathrm{ID}}$. (c) Performance by post-hoc context attribution and controlled intervention on the domain-knowledge subset.

(a) Prior on abstract patterns

	z Recall@5	Hit@5
MINA-GABT	–	$21.20_{\pm.20}$
+ population $p(z)$	30.81	$21.29_{\pm.13}$
+ context $p(z \mid C)$	**37.34**	$21.11_{\pm.33}$

(b) Oracle interventions

Oracle information	Hit@5
None	$21.20_{\pm.20}$
z^*	$36.39_{\pm.38}$
c^*	$47.97_{\pm.33}$
c^*, z^*	$56.40_{\pm.37}$
c^*, p^*	$\mathbf{77.65}_{\pm.22}$

(c) External context and controlled intervention

Performance by potentially useful external context

Attribution	Direct	MINA-GABT
Creator history (11.22%)	$10.43_{\pm.25}$	$19.82_{\pm1.32}$
Domain knowledge (40.17%)	$12.67_{\pm.56}$	$21.07_{\pm.93}$
Community culture (4.10%)	$11.24_{\pm1.84}$	$21.29_{\pm.70}$
Contemporary context (4.74%)	$14.93_{\pm1.20}$	$19.44_{\pm.60}$
Any external (60.23%)	$12.33_{\pm.34}$	$20.73_{\pm.55}$
None (39.77%)	$24.43_{\pm1.04}$	$36.81_{\pm.50}$

Controlled intervention on domain-knowledge events

Input condition	$\mathrm{Hit@5}_{B=10}$	Hit@10
MINA-GABT	$13.61_{\pm.57}$	$21.07_{\pm.93}$
+ 5 context facts	$14.51_{\pm.33}$	$21.77_{\pm.68}$
+ 1 oracle fact + 4 context facts	$\mathbf{26.12}_{\pm.19}$	$\mathbf{39.20}_{\pm.31}$

additionally investigate whether historical viewer behavior can provide such event-specific signals through retrieval. Retrieval helps when a transferable precedent exists (+3.08 pp) but hurts otherwise (-1.90 pp), yielding little aggregate improvement and suggesting that identifying transferable gap-formation processes is itself challenging. Refer to Appendix F for more details.

Beyond the Observable Context Event-specific gap formation may also depend on predictive information not contained in the observable video context C (Miyake & Norman, 1979; Garcia et al., 2020). We post-hoc attribute each test event to whether prediction could benefit from creator history, domain knowledge, community culture, contemporary context, or none of these sources. With $B = 10$, Table 5 shows that MINA-GABT reaches only 20.73% Hit@10 on events associated with external context, compared with 36.81% when no such context is identified. We further conduct a controlled oracle intervention on the domain-knowledge subset. We use Gemini-3.7-Flash with Google Search to construct a target-aware oracle fact that should help prediction. We also construct 5 context-relevant background facts without access to the original question, used as context-relevant distractions. We compare 5 target-blind background facts against 1 oracle fact mixed with 4 background facts. Table 5 shows that target-blind background facts provide little benefit, while the oracle condition improves Hit@10 by 18.13 pp, suggesting that **part of the remaining uncertainty lies in information outside the observable context**. Though not validated in this work, viewer personalization may also constitute another source of such information (Stewart & Mihalcea, 2022).

7 Conclusion

We introduced the task MINA and associated dataset MINA-Data for anticipating the information needs real viewers express at a given video moment. MINA-GABT predicts a structured gap-formation state before generating each question, using a taxonomy induced from training-set viewer behavior to guide prediction. On the main test set, it improves Hit@5 from the best direct-forecasting result of 15.53% to 21.20% and Hit@10 from 19.32% to 27.12%; the advantage persists on a held-out content category.

More broadly, **MINA reveals a gap between semantic plausibility and behavioral anticipation**. Generating plausible questions is relatively easy, but identifying the needs viewers actually express remains difficult. Oracle analyses point to event-specific gap formation and relevant information beyond the observed video as important sources of remaining uncertainty. Because the data come from one platform and evaluation measures answer coverage rather than deployed usefulness, transfer across communities and benefits to viewers remain to be tested.

AI USE STATEMENT

Generative AI tools assisted the authors in discussing and organizing the conceptual framework, locating relevant literature in psychology and cognitive science, and discussing experimental design. The authors made the final decisions about the framework and experimental design and manually verified all cited literature. LLM APIs were used in the dataset construction workflow to filter candidate comments and produce canonical information-need annotations. Generative AI also assisted with code generation; the authors inspected the resulting code and reviewed and revised the prompts used for prediction and evaluation. Generative AI assisted with exploratory analysis and interpretation of experimental results. Authors independently verified the reported results and made the final decisions regarding their interpretation and presentation.

For manuscript preparation, the authors first specified the overall outline and intended arguments, and generative AI produced an initial draft. Every sentence in the final manuscript was edited or verified by the authors. Figure 1 was generated with an AI tool, while its structure, representations, and specific content were determined by the authors and refined through multiple rounds of review to ensure that it expresses their intended meaning. The use of LLM/VLM APIs as components of the research method and evaluation is documented in the corresponding sections. The authors take responsibility for the final text, claims, code, figure, and other research artifacts.

ETHICS STATEMENT

MINA-Data is constructed from Bilibili videos and naturally occurring, time-aligned viewer comments. Its event records include original comment text, video IDs, and timestamps, but no usernames, account IDs, or pseudonymous user identifiers. Even without user identifiers, these fields may allow a comment to be located on the source platform, so we do not regard the records as fully anonymous. Hosted model APIs also process comment content during annotation and evaluation, as disclosed in the AI use statement. Whether video files can be redistributed will be decided separately in light of platform conditions and creators' rights. The human validation in Appendix A was performed by the authors; no external annotators were recruited. Our experiments evaluate moment-level patterns without intervening in viewers' experiences or modeling individual user profiles. Nonetheless, information-need anticipation could be repurposed for profiling or persuasive targeting, and the sampled Bilibili content does not represent all viewing communities.

REPRODUCIBILITY STATEMENT

Section 3.2 describes the data collection criteria and video-disjoint splits, Section 3.3 defines the evaluation criterion, and Appendix C specifies the observation window, model assignments, prediction budgets, and repeated judging. Appendix E reports the taxonomy inventory and induction statistics. Code and MINA-Data with the fixed splits and annotations will be released upon acceptance. Several experiments use hosted LLM/VLM APIs, whose outputs may vary as providers update their models.

REFERENCES

Anna Bodonhelyi, Enkeleda Thaqi, Süleyman Özdel, Efe Bozkir, and Enkelejda Kasneci. From passive watching to active learning: empowering proactive participation in digital classrooms with ai video assistant. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pp. 1–21, 2025.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. Beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*, 2022.

Dibyadip Chatterjee, Zhanzhong Pang, Fadime Sener, Yale Song, and Angela Yao. Don't pause! every prediction matters in a streaming video. *arXiv preprint arXiv:2604.24317*, 2026.

Maryam S Cheema, Sina Elahimanesh, Pooyan Fazli, and Hasti Seifi. Vidscribe: Multimodal ai for customizing audio description and question answering in online videos. In *Proceedings of the*

Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems, pp. 1–6, 2026.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024.

Ahmed Elgohary, Denis Peskov, and Jordan Lee Boyd-Graber. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5918–5924, 2019.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10826–10834, 2020.

Arthur C Graesser and Cathy L McMahen. Anomalous information triggers questions when adults solve quantitative problems and comprehend stories. *Journal of Educational Psychology*, 85(1): 136, 1993.

Yue Jiang, Jichu Li, Yang Liu, Dingkang Yang, Feng Zhou, and Quyu Kong. Danmakutppbench: A multi-modal benchmark for temporal point process modeling and understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6544–6555, 2020.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Daniel J Liebling, Paul N Bennett, and Ryen W White. Anticipatory search: using context to initiate search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 1035–1036, 2012.

George Loewenstein. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, 116(1):75, 1994.

Naomi Miyake and Donald A Norman. To ask a question, one must know enough to know what is not known. *Journal of verbal learning and verbal behavior*, 18(3):357–364, 1979.

Jose Otero and Arthur C Graesser. Preg: Elements of a model of question asking. *Cognition and instruction*, 19(2):143–175, 2001.

Arpan Phukan, Manish Gupta, and Asif Ekbal. Ecis-vqg: Generation of entity-centric information-seeking questions from videos. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14411–14436, 2024.

Arpan Phukan, Anupam Pandey, Deepjyoti Bodo, and Asif Ekbal. Videochain: A transformer-based framework for multi-hop video question generation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2246–2266, 2025.

Weihang Qin, Feng Xiao, Yuefan Fang, Zijing Tang, Shimin Kang, Zhuo Huang, and Tao He. Epimopilot: An llm-infused epistemic emotion support system to boost self-regulated asynchronous online video learning. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2026.

Ashwin Ram. A theory of questions and question asking. *Journal of the Learning Sciences*, 1(3-4): 273–318, 1991.

Bradley J Rhodes and Pattie Maes. Just-in-time information retrieval agents. *IBM Systems journal*, 39(3.4):685–704, 2000.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Hyungyu Shin, Eun-Young Ko, Joseph Jay Williams, and Juho Kim. Understanding the effect of in-video prompting on learners and instructors. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018.

Ian Stewart and Rada Mihalcea. How well do you know your audience? toward socially-aware question generation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 255–269, 2022.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026.

Weiying Wang, Jieting Chen, and Qin Jin. Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2599–2607, 2020.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. *arXiv preprint arXiv:2507.09313*, 2025.

Qunfang Wu, Yisi Sang, Shan Zhang, and Yun Huang. Danmaku vs. forum comments: understanding user participation and knowledge sharing in online videos. In *Proceedings of the 2018 ACM international conference on supporting group work*, pp. 209–218, 2018.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1666–1677. IEEE, 2021.

Ke Yang, Kevin Ros, Shankar Kumar Senthil Kumar, and ChengXiang Zhai. Jir-arena: The first benchmark dataset for just-in-time information recommendation. *arXiv preprint arXiv:2505.13550*, 2025.

Saelyne Yang, Sunghyun Park, Yunseok Jang, and Moontae Lee. Ytcommentqa: Video question answerability in instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19359–19367, 2024a.

Saelyne Yang, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. Aqua: Automated question-answering in software tutorial videos with visual anchors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2024b.

Yichi Zhang, Xin Luna Dong, Zhaojiang Lin, Andrea Madotto, Anuj Kumar, Babak Damavandi, Joyce Chai, and Seungwhan Moon. Proactive assistant dialogue generation from streaming egocentric videos. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12055–12079, 2025.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing LYU, and Xirong Li. Omnipro: A comprehensive benchmark for omni-proactive streaming video understanding. *arXiv preprint arXiv:2605.18577*, 2026.

A HUMAN VALIDATION

Dataset Validation. We conduct a human audit to verify the validity of the behavioral events and the fidelity of their canonical annotations. We randomly sample 200 events from $\mathrm{test}_{\mathrm{ID}}$, 100 events from $\mathrm{test}_{\mathrm{OOD}}$ and 200 events from train for validation. Human reviewers inspect the events and their canonical annotations, to determine (1) whether the comment expresses a genuine information-seeking intent, (2) whether its intended information need can be reliably interpreted from the observable context, and (3) whether X^* clarifies the intended information need. The resulting acceptance rates are 98.17% for information-seeking validity, 96.83% for question interpretability based on the video context, and 95.83% for canonical fidelity.

Coverage Evaluator Validation. We validate the LLM-based evaluator along two complementary dimensions: evaluation stability and agreement with human judgments. All judging is repeated three times with the same predictions, and the averaged pairwise agreement is 99.01%. We also validate the evaluator against independent human judgments using the same directional information-coverage relation in Eq. 2. Specifically, we randomly sample 100 events from $\mathrm{test}_{\mathrm{ID}}$. We examine predictions with $B = 5$ from MINA-GABT and Direct Forecasting (Gemini-3.7-Flash) for these events. For each prediction, human reviewers are shown the original question, its canonical annotation, and a predicted question at a time, and independently determine whether $\hat{Q} \succeq Q$. The averaged agreement is 98.2%, showing the stability and effectiveness of the evaluator.

Plausibility Auditor Validation. We further validate the VLM-based plausibility auditor used in Section 6 against independent human judgments. We randomly sample 50 events from $\mathrm{test}_{\mathrm{ID}}$. For each event, all the predictions judged by the plausibility auditor are reviewed, including 10 predictions from Direct Forecasting, 10 predictions from MINA-GABT, 5 cross-video predictions from Direct Forecasting and 5 cross-video predictions from MINA-GABT. For each event-prediction pair, human reviewers are shown the observable video context C and one predicted question, without access to the target viewer question, and independently determine whether the prediction expresses a semantically plausible information need given C. The auditor agrees with human judgments on 96.27% of the sampled pairs (95.5% on matched-context predictions and 97.8% on cross-video negative controls), supporting its use for the plausibility analysis.

Prediction Specificity Audit. As our information-coverage criterion is asymmetric, we additionally examine whether evaluated hits can be obtained through overly broad or under-specified predictions. We randomly sample 200 consensus hit predictions (i.e. predictions that are judged as hits in all three runs) from Direct Forecasting and MINA-GABT on $\mathrm{test}_{\mathrm{ID}}$. Human reviewers inspect each predicted question together with the corresponding ground-truth question and canonical annotation, and categorize it as a specific coherent information need, a legitimately broader information need, a compound prediction, or a generic catch-all prediction. Among all sampled predictions, 95.5% are specific or legitimately broader information needs, while 3.5% are compound and 1% are generic catch-all predictions. These results indicate that the observed coverage gains are not primarily driven by overly broad predictions.

B COGNITIVE MOTIVATION FOR THE GAP-FORMATION STATE

Our representation separates what in the context attracts attention, what information the viewer seeks, and how that need is expressed as a question. This separation is motivated by several related accounts of question asking, although none specifies our tuple $S = (c, p, g)$. Ram (1991) shows that an inadequate domain model in a reasoning situation can create a knowledge goal: the reasoner identifies needed knowledge and formulates a question to acquire it. This motivates representing an information need before its linguistic realization as Q. Graesser & McMahen (1993) likewise discuss anomaly detection and question articulation as distinct, interacting components of question asking.

For the context-to-gap transition, PREG explains question asking during expository-text reading through discrepancies between the reader's representation of the text and prior world knowledge, with pragmatics and metacognition mediating question production (Otero & Graesser, 2001). In our video setting, c denotes an observable cue that may become relevant to a viewer; p records the

inferred process by which that cue points to a particular missing piece of information. This is an operational analogy to discrepancy-driven question formation, not an implementation of PREG's production rules. Experimental evidence also supports selective cue effects. Graesser & McMahen (1993) varied quantitative problems and stories by deleting critical information, adding contradictions, or adding irrelevant information. Some anomalies elicited more questions than the original materials, and some questions directly addressed those anomalies; deletions elicited more questions than the other altered versions. Thus, the evidence motivates identifying question-relevant cues rather than treating every observed event as equally likely to elicit a question.

Loewenstein (1994) proposes an information-gap account of specific state curiosity, in which curiosity arises when a person perceives a gap in knowledge or understanding. This provides motivation for explicitly representing the target of an information need as g, without assuming that every viewer question is driven by curiosity. Prior knowledge can also affect whether a question is asked: Miyake & Norman (1979) found that novice and trained learners asked different numbers of questions depending on the difficulty of the learning material. This cautions against treating question emergence as a function of video content alone.

These studies concern reading, learning, problem solving, and curiosity rather than naturalistic video comments. They motivate our computational factorization $C \to c \xrightarrow{p} g \to Q$, but do not establish it as a literal or fixed temporal sequence in viewers' minds. In particular, p is our explanatory description of a cue-to-gap relation, not a directly observed cognitive mechanism. The empirical value of the structured state for MINA is examined by the "Intermediate S only" versus "Free-form need only" comparison in Table 3.

C Experimental Setup

Metrics. We evaluate MINA under two prediction budgets, $B = 5$ and $B = 10$. For each budget B, the model is independently prompted to produce $\hat{\mathcal{Q}}^{(B)} = \left(\hat{Q}_1^{(B)}, \ldots, \hat{Q}_B^{(B)}\right)$. As the budget can affect the generated predictions, predictions under different budgets are generated and evaluated independently. For event i, we define the first covered rank under budget B as

$$r_i^{(B)} = \min\left\{k \in \{1, \ldots, B\} \;\middle|\; \hat{Q}_{i,k}^{(B)} \succeq Q_i\right\}, \tag{7}$$

with $r_i^{(B)} = \infty$ if no prediction hits. For any $K \leq B$, we define the budget-conditioned Hit@K and budget-conditioned mean reciprocal rank as:

$$\text{Hit@}K_B = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[r_i^{(B)} \leq K\right], \tag{8}$$

$$\text{MRR@}B = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i^{(B)}}, \tag{9}$$

where $1/r_i^{(B)} = 0$ when $r_i^{(B)} = \infty$. We typically report Hit@5 := Hit@$5_{B=5}$ and Hit@10 := Hit@$10_{B=10}$. Predictions under each budget are generated once. To provide a robust evaluation, we repeat the judging process and report the mean and standard deviation across evaluator runs.

Evaluation splits and prediction budgets. We follow the evaluation protocol in Section 3.3. The primary evaluation uses test_{ID} with prediction budgets $B = 5$ and $B = 10$; category-held-out generalization uses test_{OOD} with $B = 5$. Each budget is a separate generation condition: we prompt each method to produce exactly B ranked questions and generate the $B = 5$ and $B = 10$ lists independently. The main results report Hit@5, MRR@5, Hit@10, and MRR@10, respectively computed from the matching budget.

Observable context. Unless otherwise specified, a video-based method receives the video title and description together with the 10-second trailing video window ending at the prediction timestamp t. Video is supplied at 1 fps. For frame-based inputs, we additionally provide a timestamped audio transcript restricted to the same observation window; transcript content after t is excluded. The

ASR and its word-level timestamp are generate by qwen3-asr-1.7b and Qwen3-ForcedAligner-0.6B (Shi et al., 2026). The video-history sensitivity experiment changes the trailing-window length while holding the prediction timestamp and metadata fixed. At test time, prediction methods do not receive the observed viewer question or its canonical annotation.

Methods and model configuration. Direct Forecasting asks a single model to predict the ranked questions directly from the observable context. Table 3 compares Qwen3.8-Flash, GLM-5.3-Flash, Gemini-3.7-Flash, Gemini-3.8-Flash, Grok-4.7, GPT-5.6-Luna, GPT-5.6-Terra, GPT-5.6-Sol, and GPT-6-Astra under this setup. Unless otherwise specified, MINA-GABT uses Gemini-3.7-Flash to anticipate gap-formation states and GPT-5.6-Luna to realize them as questions. The "Intermediate S only" variant omits taxonomy conditioning, while "Taxonomy $\mathcal{T}$ only" predicts questions directly with the taxonomy but without the intermediate state. "Free-form need only" uses the same prediction and realization models as "Intermediate S only", but replaces the structured state $S = (c, p, g)$ with an unconstrained free-form information-need description. We enable thinking for the prediction models and use their default thinking effort. Taxonomy induction uses GPT-5.6-Sol for the curator and analyst and GPT-5.6-Luna for auditing; the taxonomy is committed and fixed before test evaluation. All dataset-dependent method development, including taxonomy induction and prompt selection, is confined to `train`, with video-disjoint internal development and validation subsets.

Judging and aggregation. We generate each prediction list once and hold it fixed across evaluator repeats. The question-hit evaluator receives the original viewer question, its canonical annotation X^*, and the ranked predictions, and judges each predicted question independently using the asymmetric answer-coverage criterion in Section 3.3. We use GPT-5.6-Luna with high thinking effort for three independent judging runs. For each run, we compute Hit@K and MRR from the rank of the first covered prediction; tables report the mean and sample standard deviation of these metrics across the three runs.

D Taxonomy Induction Action Space

At each step, the taxonomy induction agent selects one of the following actions exposed by the harness:

- `analyze` asks a stateless analyst to extract reusable gap-formation patterns from all or selected evidence, without changing the taxonomy.
- `audit` asks a stateless auditor to assess the current taxonomy on all or selected evidence, returning coverage and boundary diagnostics together with records of uncovered or problematic cases.
- `inspect_evidence` reveals the oracle gap information for selected evidence references.
- `inspect_leaves` retrieves selected taxonomy patterns and, when an audit is available, their support counts and example cases from that audit.
- `initialize_taxonomy` creates the initial taxonomy.
- `modify_taxonomy` applies a small, evidence-supported revision to the current taxonomy by adding, redefining, deleting, merging, or splitting patterns.
- `validate_dev` validates the current taxonomy on a protected, video-disjoint validation set and returns only aggregate statistics, without revealing individual examples, references, or judgments.
- `finish` ends the agent loop and commits the taxonomy. It must be done after the current version has been audited and validated on the protected set.

E Final Behavior-derived Taxonomy

The committed taxonomy is a flat list of 37 non-exclusive patterns. Table 7 gives their IDs and concise English descriptions. Examples from the final induction audit illustrate the distinction between patterns: an uncredited background song prompts a source query (`unattributed_artifact_provenance`); omitted preparation before food is eaten prompts

Table 6: Committed taxonomy and audit statistics. Coverage means that at least one leaf accurately describes the oracle gap's cue-to-gap process. The test result is a separate post-hoc analysis and was unavailable during induction.

Statistic	Value
Patterns	37
Coverage on development set	96.4%
test_{ID} coverage (post hoc)	92.24%

recovery of prior steps (`process_gap_completion`); and an observed high-speed effect prompts a question about a further speed threshold (`boundary_condition_stress`).

Table 7: Index of all committed taxonomy patterns. Descriptions summarize the definitions.

Leaf ID	Mechanism summarized in English
`unlabeled_referent_grounding`	A visible referent lacks a basic name or identity.
`incomplete_signal_recovery`	Weak or occluded perceptual evidence leaves content or identity unresolved.
`deliberate_concealment_probe`	An intentional mask or mute makes hidden content or its motive salient.
`candidate_correspondence_check`	A perceptual cue suggests a specific known candidate whose identity must be checked.
`unattributed_artifact_provenance`	A recognizable work or product lacks attribution or source.
`utterance_text_recovery`	Speech or singing is apparent but its literal words are unavailable.
`symbol_semantic_grounding`	A visible or audible symbol or term lacks a contextual meaning.
`underspecified_reference_completion`	A pronoun, generic phrase, or ellipsis leaves a referent or value unfilled.
`diagnostic_feature_inference`	Observable features suggest an unobserved attribute or state.
`schema_slot_absence`	A familiar structure implies a missing member, component, or role.
`explicit_remainder_closure`	A count, fraction, or near-complete list exposes an explicit remainder.
`overview_focal_projection`	An aggregate view invites retrieval of one entity's local value or position.
`sample_scope_projection`	A few examples invite questions about the wider set or range.
`process_gap_completion`	A partial procedure leaves preceding, intermediate, or paired steps unseen.
`outcome_backchain`	A visible result invites reconstruction of its unshown causal history.
`open_loop_closure`	An initiated event, promise, or task has no observed resolution.
`timeline_extension`	A temporal anchor invites an origin, precedent, or longer history.
`actionability_path_completion`	An activated practical goal lacks a route, channel, or next step.
`baseline_deviation_abduction`	A departure from an expected baseline invites explanation of its cause.
`evidence_conflict_reconciliation`	Incompatible cues invite a reconciled account of the true state.
`agentive_choice_rationale`	A salient deliberate choice invites its goal, constraint, or tradeoff.
`functional_role_inference`	An unexplained part or action invites its purpose in a larger system.
`implementation_mechanism_reconstruction`	An observed result leaves the internal process that produced it unknown.
`production_path_disambiguation`	One audiovisual appearance admits competing production processes.

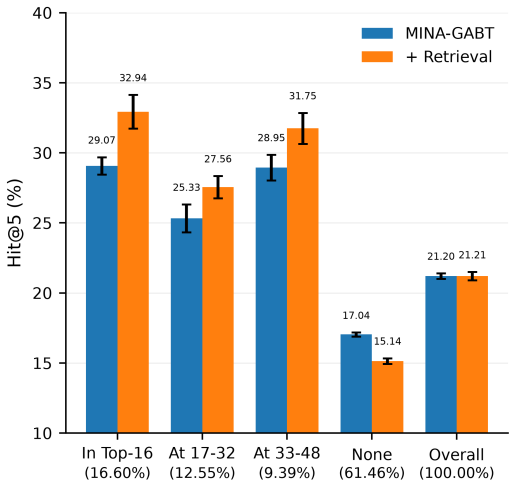

Figure 4: Results by the rank intervals of the first transferable retrieval entry.

Leaf ID	Mechanism summarized in English
`relation_hypothesis_completion`	Co-occurrence or interaction suggests a relationship to verify.
`intertextual_influence_check`	A specific resemblance invites a test of homage, copying, or influence.
`pragmatic_implicature_recovery`	Literal speech is clear but its intended implication remains unresolved.
`claim_support_audit`	A strong claim lacks independent support or rests on a doubtful premise.
`claim_measurement_refinement`	A coarse or qualitative claim invites a precise measure or breakdown.
`external_norm_retrieval`	A concrete case invokes an unstated formal rule or standard.
`given_rule_instance_test`	A stated rule invites judgment of whether a case satisfies it.
`boundary_condition_stress`	A demonstrated capability invites testing under more extreme conditions.
`counterfactual_variant_test`	A known plan invites an explicit substitution or intervention.
`prospective_risk_projection`	An observed hazard invites assessment of a not-yet-observed harm.
`remedy_search`	A visible cost or defect invites a mitigation or repair.
`comparison_dimension_resolution`	Parallel cases invite identification of their difference or advantage.
`practical_decision_variable_completion`	An adoption decision lacks a needed price, specification, or burden.

F Can Historical Viewer Behavior Provide the Missing Signal?

Existing analyses highlight the importance of context-specific instantiation. We therefore ask whether retrieval from historical viewer behavior can provide analogous signals. We construct a retrieval bank from training events annotated with S^* and, for each test event, retrieve by cue similarity within each compatible taxonomy pattern. Retrieval entries are ordered by similarity scores, truncated with a fixed budget (48), and provided to the predictor without scores. We also examine, post hoc, whether the formation process p^* of each retrieved entry is transferable to the current test event, and analyze Hit@5 by the rank intervals of the first transferable entry. Results are shown in Figure 4. Retrieval from historical behavior improves Hit@5 by 3.08 pp on events with at least one transferable precedent, while causing a 1.90 pp degradation otherwise, explaining its near-zero aggregate gain. It indicates that historical viewer behavior contains complementary event-level information, but exploiting it requires identifying whether a gap-formation process p can transfer to the current event.