REELS: Large Language Models See Every Video but Cannot Read the Room

Anonymous ACL submission

Abstract

Generative search is increasingly used to address open-ended information needs in real-world user-facing scenarios. Unlike factoid question answering, synthetic multi-hop tasks, or datasets with expert-written reference answers, many everyday search needs require models to aggregate experiences, preferences, and competing viewpoints from large-scale user-generated content. In this setting, answers are corpus-dependent rather than canonical: they should reflect the distribution of viewpoints expressed in the collected corpus. To enable systematic study, we introduce REELS, a benchmark built from real videos on a major user-generated video platform, comprising 2,710 videos spanning 67 topics across 7 domains, with 662 queries covering four representative task types. Each query is paired with structured, corpus-grounded reference answers that include normalized viewpoints, evidence snippets, and source attribution. We evaluate both open-source and proprietary models under a unified agentic generative search framework. Our experiments identify diverse evidence retrieval and faithful cross-source viewpoint aggregation as major bottlenecks. Even the best-performing models in our evaluation still exhibit substantial room for improvement in multi-video evidence synthesis.

1 Introduction

Generative search systems are increasingly used for open-ended information needs, supported by retrieval-augmented generation and browser-assisted question answering (Lewis et al., 2020a; Nakano et al., 2021). Many real user queries, however, are opinion-seeking rather than fact-seeking, asking what people experience, prefer, complain about, or recommend across diverse sources. Since such queries lack a single canonical answer, a useful response must synthesize the recurring, minority, and conflicting viewpoints expressed in the available evidence, rather than return a single asserted fact. We refer to this setting as *corpus-grounded opinion aggregation*, which targets a faithful synthesis of the corpus-level opinion landscape rather than a universal truth. This problem becomes especially challenging on user-generated video platforms, where opinions are often embedded in long, noisy, and multimodal content, including speech, on-screen text, visual demonstrations, and usage scenes. Relevant evidence may be scattered across many videos, while semantically similar viewpoints may be expressed with different wording, examples, or visual cues. Therefore, a generative search system must go beyond understanding individual videos: it must retrieve diverse sources, filter noisy candidates, identify grounded evidence, normalize similar viewpoints, preserve disagreements, and aggregate evidence into a faithful corpus-level summary.

Existing benchmarks evaluate adjacent capabilities but fall short of fully capturing this intricate paradigm. Agentic search benchmarks test retrieval and tool use, yet they predominantly target verifiable facts or expert-style reports rather than corpus-dependent opinion distributions (Mialon et al., 2023; Wei et al., 2025). Video comprehension benchmarks focus almost exclusively on single or long videos, ignoring cross-video retrieval and distribution-level synthesis (Fu et al., 2025; Wu et al., 2024). While opinion mining and review summarization share a similar spirit, they remain text-centric and typically assume that relevant input documents have already been collected, which makes it difficult to diagnose failures in evidence recall, source retention, multimodal grounding, or viewpoint aggregation (Hu and Liu, 2004; Angelidis and Lapata, 2018a; Brazinskas et al., 2020a).

We introduce REELS, a benchmark for corpus-grounded multimodal opinion aggregation over real user-generated videos. REELS contains 2,710 videos from a major user-generated video plat-

form, covering 67 topics across 7 domains and 662 opinion-seeking queries. The benchmark includes four representative task types spanning entity discovery, factor induction, comparative judgment, and proposition stance aggregation. Rather than relying on unstable live web search, REELS implements a controlled simulation retriever over a fixed corpus to preserve an interactive, multi-step search environment. The core of REELS is its structured, source-grounded reference format containing normalized viewpoint labels, concise summaries, supporting evidence snippets, and unique source counts.

Building on this structured reference format, REELS separates process-level evidence acquisition from content-level opinion aggregation, which makes failures observable beyond end-to-end answer quality. The process evaluation tracks whether models can recall, screen, and ground relevant videos, whereas the content evaluation assesses viewpoint recovery, source grounding, and distributional fidelity in the final structured answer. We evaluate representative large language models (LLMs) and multimodal large language models (MLLMs) under a unified agentic workflow, finding that current systems remain limited by recall-preserving evidence acquisition and faithful distribution-aware synthesis: they often miss long-tail or scattered evidence, improve precision by discarding relevant sources during screening, and misestimate source support even when relevant evidence is available. Video-aware input helps on visually grounded queries, but its gains do not consistently transfer to the full benchmark, suggesting that visual information functions as a conditional evidence source rather than a universal augmentation.

Our primary contributions are clustered into three aspects: **(1)** We formalize corpus-grounded multimodal opinion aggregation as a realistic generative search task and release REELS, a benchmark built from real user-generated videos across diverse domains; **(2)** We design structured, distribution-aware reference answers alongside a diagnostic evaluation protocol that decouples evidence acquisition from final synthesis; **(3)** We evaluate representative LLMs and MLLMs under text-only and video-aware settings, uncovering critical bottlenecks in evidence recall, source retention, viewpoint normalization, and visual utilization.

2 Related Work

Web-Grounded Generation and Deep Research Benchmarks. Web-grounded and retrieval-augmented language models have become central approaches for answering information-seeking queries with external evidence (Lewis et al., 2020b; Nakano et al., 2021; Asai et al., 2024). Recent benchmarks on agentic and deep-research settings further test whether models can use tools, browse the web, and carry out multi-step information seeking (Mialon et al., 2024; Zhou et al., 2023; Wei et al., 2025; Du et al., 2025). These benchmarks have advanced the evaluation of retrieval, tool use, and complex reasoning, but they are largely framed around factual questions, task completion, or expert-oriented report writing. Everyday opinion-seeking search poses a different problem: the answer is often not a single canonical fact, but a set of recurring, conflicting, and sometimes long-tail viewpoints grounded in the available sources.

Opinion Mining and Opinion Summarization. Opinion mining, sentiment analysis, and aspect-based sentiment analysis provide a long-standing foundation for modeling subjective user-generated content, including the extraction of opinions, sentiments, and opinion targets (Amplayo et al., 2022; Pang and Lee, 2007; Pontiki et al., 2014). Building on this line of work, opinion summarization studies how to aggregate subjective signals from collections of reviews or social media posts, often by identifying salient aspects, estimating sentiment polarity, and generating compact summaries of user opinions (Angelidis and Lapata, 2018b; Chu and Liu, 2019; Brazinskas et al., 2020b; Amplayo et al., 2021). These tasks are relevant to opinion-seeking search because they address how user attitudes and preferences can be represented once a set of opinion-bearing documents is available. However, they typically assume a pre-collected textual input set, and thus do not evaluate whether a system can search for, filter, and select relevant evidence from a larger corpus. Their outputs also differ: prior work typically yields sentiment labels, aspect-level summaries, or free-form summaries, whereas opinion-seeking search requires structured summaries that organize recurring viewpoints and ground them in their source evidence.

Multimodal and Video Understanding Benchmarks. A growing body of benchmarks evaluates

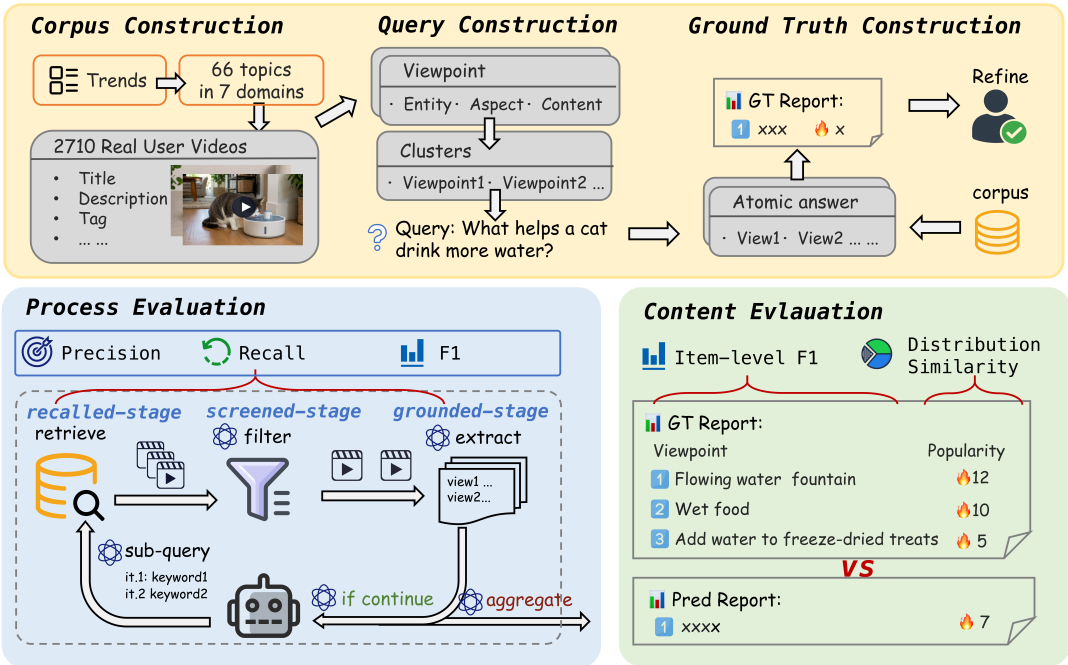

Figure 1: Overview of the REELS benchmark construction and evaluation framework. REELS constructs queries and structured reference reports from recurring viewpoints grounded in a corpus of user-generated videos. Evaluation is conducted at both the process level, measuring evidence discovery across recalled, screened, and grounded stages, and the content level, measuring viewpoint recovery and source-count distribution fidelity.

multimodal models on video question answering, temporal reasoning, event understanding, and long-video comprehension (Li et al., 2024; Zhou et al., 2025; Wu et al., 2024). These benchmarks are essential for measuring whether models can understand visual, auditory, and temporal information in videos. However, they typically focus on individual or pre-specified videos, and the expected answer is often a factual response, a multiple-choice option, or a short explanation.

3 REELS

3.1 Task Definition

REELS studies corpus-grounded opinion aggregation over user-generated videos. Given an opinion-seeking query q and a video corpus $\mathcal{V}$, the system produces a structured answer that summarizes the viewpoints expressed in the corpus. Each answer contains a set of viewpoint items and an overall summary:

$$\mathcal{A}_q = (\{a_k\}_{k=1}^{K}, g_q) \quad (1)$$

$$a_k = (\ell_k, m_k, E_k, B_k, c_k), \qquad c_k = |B_k|. \quad (2)$$

Here, ℓ_k is a normalized viewpoint label, m_k is a concise viewpoint summary, E_k contains supporting evidence snippets, B_k is the set of source video identifiers, and c_k is the number of unique supporting videos.

3.2 Task Taxonomy

We categorize real-world user-opinion search needs into four task types, which correspond to four common user intents in generative search: discovering entities, inducing factors, comparing options, and aggregating stances. More complex user queries can often be decomposed into one or more of these basic types.

Entity Discovery and Ranking. EDR asks the model to identify a set of concrete entities and rank them by source count. It evaluates whether the model can recognize entities, normalize aliases or informal mentions, aggregate multi-source evidence, and estimate the relative prominence of entities. For example, a user may ask, "Which Japanese restaurants in Shanghai are most frequently recommended?"

Factor Induction and Aggregation. FIA asks the model to infer abstract factors, reasons, methods, or evaluation dimensions from multiple user opinions. It evaluates abstraction, semantic normalization, viewpoint clustering, and the ability to preserve both dominant and long-tail concerns. For

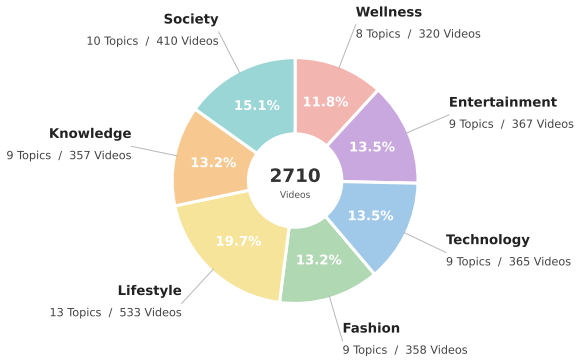

Figure 2: Domain and topic distribution of REELS. Each label reports the number of topics and videos in the corresponding domain.

Category	Num	Pct.
Query Type		
Key Factor Extraction	439	66.3%
Entity Listing	96	14.5%
Stance Judgment	82	12.4%
Multi-Entity Comparison	45	6.8%
Video Duration		
<3 min	371	13.7%
3–10 min	1,124	41.5%
10–30 min	977	36.1%
>30 min	238	8.8%

Table 1: Statistics of query types and video durations

example, a user may ask, "What factors do users care about most when choosing an air purifier?"

Multi-Entity Comparative Judgment. MECJ asks the model to compare two or more specified or implied entities under a query-specific aspect or preference dimension. It evaluates balanced retrieval across entities, entity-level evidence aggregation, contrastive reasoning, and comparative summarization. For example, a user may ask, "Are AirPods Pro or Sony WF-1000XM5 more widely praised for noise cancellation?"

Proposition Stance Aggregation. PSA asks the model to aggregate user opinions toward a single proposition, claim, or decision statement and estimate the overall stance distribution. It evaluates stance identification, disagreement modeling, evidence aggregation, and the ability to distinguish support, opposition, mixed views, and neutral attitudes. For example, a user may ask, "Does drinking sugar-free drinks for a long time affect health?"

3.3 Dataset Statistics

REELS contains 2,710 user-generated videos, 662 opinion-seeking queries, and 67 topics across seven domains. Figure 2 presents the domain and topic distribution, while Table 1 summarizes the query-type and video-duration distributions. Overall, the benchmark covers diverse everyday domains, multiple opinion-seeking task types, and videos of varying lengths, supporting evaluation of multi-source opinion aggregation over realistic user-generated video content.

3.4 Video Corpus Construction.

We build our corpus from Bilibili, a large-scale Chinese user-generated video platform. To obtain topic seeds that reflect broad user interests, we first collect candidate topics from Bilibili trending lists and consolidate noisy or duplicated entries through semantic clustering and manual inspection. For each topic, we retrieve videos with keyword-based search using aliases, abbreviations, aspect-specific keywords, and colloquial expressions. After removing duplicate or inaccessible videos, we record platform metadata and construct retrievable resource paths.

3.5 Query Construction

We construct queries from recurring viewpoints in the video corpus. This viewpoint-driven pipeline ensures that each query reflects a real user concern and is grounded in multi-video evidence.

Viewpoint Extraction. We use Gemini-3-Pro to extract subjective viewpoints from each video, focusing on opinions, evaluations, experience-based feedback, preferences, complaints, and recommendations, while filtering out directly verifiable facts such as specifications or static attributes. Each viewpoint is normalized as $u_j = (e_j, a_j, v_j, r_j, d_j)$, where e_j is the target entity or topic, a_j is the aspect, v_j is the subjective viewpoint, r_j is the evidence snippet, and d_j is the source video. This converts noisy video content into structured, evidence-grounded viewpoint units.

Viewpoint Clustering. We then cluster extracted viewpoints to identify recurring concerns across videos. Specifically, we encode each viewpoint with Qwen3-Embedding-8B, construct a nearest-neighbor similarity graph, and apply Leiden community detection to group semantically related viewpoints. Each cluster corresponds to a recurring

concern, such as battery life, service quality, or driving assistance experience.

Query Generation. For each viewpoint cluster supported by multiple videos, we prompt Gemini-3-Pro to generate a natural user query that could appear in a video platform, forum or search scenario. This pipeline derives queries from real user opinions rather than annotator assumptions. Since each query is grounded in a multi-video viewpoint cluster, the resulting query set is realistic, diverse, traceable, answerable, and suitable for evaluating multi-source user-opinion aggregation.

3.6 Ground-Truth Construction

For each query, Gemini-3-Pro independently annotates relevant videos by identifying query-relevant evidence and extracting grounded atomic answers. We then merge semantically equivalent atomic answers across videos into structured reference items, each containing a normalized viewpoint label, a concise viewpoint summary, supporting evidence snippets, associated source identifiers, and the number of supporting sources. The final reference answer consists of these structured reference items together with a summary of the aggregated opinion distribution. Finally, trained annotators verify aggregated references for relevance, evidence faithfulness, completeness, and merging quality, while supplementing missing answers when needed.

4 Experiments

4.1 Experimental Setup

Models. We evaluate both proprietary and open-source models under a unified agentic search workflow. The evaluated models include DeepSeek-V4-Pro, Gemini3.1-Pro, GPT-5.2, Qwen3.5-35B-A3B, Qwen3.5-397B-A17B, and Qwen3.5-Omni-Plus where applicable. All models are evaluated with thinking mode enabled and temperature set to 1. We consider two input settings. In *Transcript-Only* mode, the model analyzes videos by invoking an ASR tool to convert them into textual transcripts. In *Video-Aware* mode, the model can additionally use visual information. The same corpus, tool interface, and evaluation protocol are used across models to ensure comparability.

Agentic Search Workflow. We instantiate a unified agentic search workflow over the REELS corpus. The workflow consists of a controller LLM and five tools: keyword generation, video retrieval, ASR transcription, viewpoint extraction, and viewpoint aggregation. The controller iteratively invokes these tools, observes intermediate results, and decides how to proceed. The workflow contains five stages. First, in *Dynamic Query Generation and Retrieval*, the controller generates search queries and retrieves candidate videos from the controlled corpus. Second, in *Metadata Screening*, it performs lightweight relevance filtering based on video metadata. Third, in *Content-level Evidence Identification*, the retained videos are analyzed to identify query-relevant atomic viewpoints and supporting evidence. Fourth, in *State Evaluation and Iterative Routing*, the controller decides whether to explore new search directions, exploit existing queries, or terminate the search. Finally, in *Global Viewpoint Aggregation*, all grounded atomic viewpoints are normalized, merged across sources, and converted into the final structured answer.

Evaluation Metrics. We evaluate each model from both process and content perspectives. Process evaluation measures whether the agent can discover, preserve, and ground relevant evidence throughout the workflow. We report Precision, Recall, and F1 at three observable checkpoints: *Recalled*, the candidate videos returned by retrieval; *Screened*, the videos retained after metadata-based filtering; and *Grounded*, the videos verified through content-level analysis as containing query-relevant evidence. Content evaluation measures whether the final answer faithfully recovers the reference viewpoints and their distribution. We report item-level F1 for all task types. To evaluate distribution fidelity, we use JS-based similarity for Entity Discovery and Ranking, Factor Induction and Aggregation, and Multi-Entity Comparative Judgment. For Proposition Stance Aggregation, we use a Wasserstein-1 (W1)-based similarity because stance labels lie on an ordered semantic axis, where confusing nearby stances should be penalized less than confusing opposing stances. The overall content score is computed as the macro-average across task-level content metrics.

4.2 Main Results

Tables 2 and 3 evaluate two REELS stages: evidence acquisition and opinion synthesis. The former assesses whether models can retrieve, filter, and ground relevant video evidence; the latter assesses whether they can synthesize it into faithful, distribution-aware opinion reports.

Model	Recalled			Screened			Grounded		
	P	R	F1	P	R	F1	P	R	F1
Transcript-Only Mode									
Qwen3.5-35B-A3B	**0.369**	0.573	0.379	0.504$_{+0.135}$	0.549$_{-0.024}$	0.455$_{+0.076}$	0.760$_{+0.256}$	0.482$_{-0.068}$	0.556$_{+0.101}$
Qwen3.5-397B-A17B	0.299	0.719	0.378	0.441$_{+0.143}$	0.665$_{-0.054}$	0.462$_{+0.084}$	0.725$_{+0.284}$	**0.509**$_{-0.155}$	0.557$_{+0.095}$
Gemini3.1-Pro	0.368	0.624	**0.411**	0.565$_{+0.197}$	0.507$_{-0.117}$	0.467$_{+0.056}$	**0.851**$_{+0.286}$	0.437$_{-0.070}$	0.549$_{+0.081}$
DeepSeek-V4-Pro	0.246	0.609	0.324	**0.592**$_{+0.346}$	0.319$_{-0.290}$	0.338$_{+0.014}$	0.665$_{+0.073}$	0.194$_{-0.126}$	0.270$_{-0.068}$
GPT-5.2	0.325	0.637	0.381	0.340$_{+0.015}$	0.621$_{-0.015}$	0.388$_{+0.007}$	0.827$_{+0.486}$	0.458$_{-0.163}$	**0.562**$_{+0.174}$
Video-Aware Mode									
Qwen3.5-35B-A3B	0.318	0.675	0.371	0.463$_{+0.145}$	0.663$_{-0.012}$	0.482$_{+0.110}$	0.627$_{+0.165}$	0.255$_{-0.408}$	0.339$_{-0.143}$
Qwen3.5-397B-A17B	0.274	**0.753**	0.356	0.468$_{+0.194}$	**0.681**$_{-0.072}$	**0.508**$_{+0.152}$	0.605$_{+0.137}$	0.292$_{-0.389}$	0.364$_{-0.144}$
Qwen3.5-Omni-Plus	0.342	0.563	0.307	0.439$_{+0.097}$	0.545$_{-0.019}$	0.398$_{+0.091}$	0.615$_{+0.176}$	0.413$_{-0.132}$	0.450$_{+0.052}$
Gemini3.1-Pro	0.341	0.638	0.386	0.539$_{+0.198}$	0.597$_{-0.041}$	0.504$_{+0.118}$	0.804$_{+0.265}$	0.374$_{-0.223}$	0.491$_{-0.013}$

Table 2: Overall Recalled, Screened, and Grounded performance across models.

Initial retrieval is noisy and incomplete. Even the best recalled-stage recall remains below 0.76, indicating that a substantial portion of relevant videos is missing from the candidate pool. Meanwhile, recalled-stage precision is consistently low across models, suggesting that models introduce many irrelevant videos while still failing to fully cover the relevant evidence. This reveals a core challenge for open-ended opinion aggregation: models need not only broader search, but also more effective query expansion and coverage-oriented evidence acquisition.

Filtering improves precision but sacrifices coverage. Across models, precision generally increases from the Recalled stage to the Screened and Grounded stages, showing that models can remove noise and identify more reliable evidence. However, recall drops substantially in the same process. For example, Gemini3.1-Pro improves its grounded precision to 0.851 under the Transcript-Only setting, but its recall decreases from 0.624 at Recalled to 0.437 at Grounded. This pattern suggests that the main process-level bottleneck is not simply noise filtering, but recall-preserving evidence selection: models become more selective, yet discard too much truly relevant evidence.

Open-ended opinion discovery remains difficult. Gemini3.1-Pro achieves the best overall content score under the Transcript-Only setting, while Qwen3.5-35B-A3B is the second best. However, task-level results show that models perform better on more constrained tasks such as Multi-Entity Comparative and Stance Aggregation than on open-ended tasks such as Entity Discovery and Factor Induction. This indicates that current models are relatively effective when the label space is fixed, but still struggle to discover entities, merge aliases, identify long-tail viewpoints, and recover support distributions.

Evidence quality does not fully determine report quality. Process metrics and content metrics are not perfectly aligned. For instance, GPT-5.2 achieves the best Grounded F1 under the Transcript-Only setting, but does not obtain the best content score. Similarly, Qwen3.5-397B-A17B has strong recall in the evidence pipeline, yet its final content quality is lower than that of Qwen3.5-35B-A3B. These mismatches show that evidence acquisition is necessary but not sufficient. High-quality REELS reports require normalizing semantically equivalent viewpoints, merging duplicates, aggregating source-level support, estimating the viewpoint distribution, and synthesizing the results into a coherent structured report.

Effect of visual access. Tables 2 and 3 show that Video-Aware mode generally underperforms Transcript-Only mode, especially in grounded recall and final content quality. A plausible explanation is that many user viewpoints are explicitly expressed in speech and can already be recovered from transcripts, while additional video frames increase context length and introduce redundant or distracting signals, which can complicate evidence verification and adversely affect the controller's decisions about when to explore, exploit, or terminate. However this full-benchmark result does not imply

Model	Entity Discovery		Factor Induction		Multi-Entity Comp.		Stance Aggregation		Overall
	F1	JS-Sim	F1	JS-Sim	F1	JS-Sim	F1	W1-Sim	Avg.
Transcript-Only Mode									
Qwen3.5-35B-A3B	0.5886	0.6245	0.7297	0.7757	0.7421	0.7974	0.7992	0.7520	0.7261
Qwen3.5-397B-A17B	0.5144	0.5558	0.6961	0.7405	0.6843	0.7448	0.8150	0.7823	0.6916
Gemini3.1-Pro	**0.5969**	**0.6434**	**0.7423**	0.7906	**0.8031**	**0.8494**	0.8329	0.8137	**0.7590**
DeepSeek-V4-Pro	0.3997	0.4338	0.4804	0.5279	0.5745	0.6102	0.4688	0.4326	0.4910
GPT-5.2	0.5494	0.6037	0.6802	0.7331	0.7065	0.7753	0.7949	0.7479	0.6989
Video-Aware Mode									
Qwen3.5-35B-A3B	0.4242	0.4542	0.6858	0.7347	0.5836	0.6573	0.7038	0.6603	0.6130
Qwen3.5-397B-A17B	0.3959	0.4252	0.6718	0.7197	0.5488	0.6215	0.7357	0.6779	0.5996
Qwen3.5-Omni-Plus	0.3165	0.3340	0.6747	0.7169	0.6267	0.7186	**0.9333**	**0.8306**	0.6439
Gemini3.1-Pro	0.5693	0.6106	0.7404	**0.7939**	0.7911	0.8391	0.8262	0.7931	0.7455

Table 3: Content Quality Evaluation. F1 and distribution similarity per task type.

that visual information is unnecessary. Rather, it suggests that the benefit of video modality is conditional and is likely concentrated on queries whose key evidence is visually grounded. We further analyze this phenomenon in Section 4.3.

4.3 Analysis

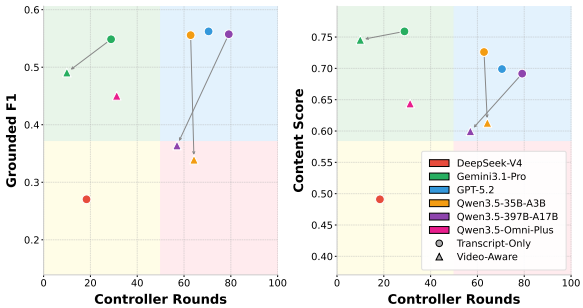

Figure 3: Cost-effectiveness of agentic search across model configurations. Upper-left points indicate better cost-effectiveness.

Cost-effectiveness of Agentic Search. We further examine the cost-effectiveness of agentic search in Figure 3. Controller rounds approximate the reasoning and orchestration cost, while Content Score measures final opinion aggregation quality and Grounded F1 captures process-level evidence grounding. The results show that higher controller cost does not consistently yield better performance. Gemini3.1-Pro lies near the favorable efficiency–quality frontier, achieving strong content quality with substantially fewer steps than GPT-5.2 and the Qwen models. By contrast, GPT-5.2 and the Qwen transcript-only variants attain competitive Grounded F1 but require many more controller rounds, indicating higher reasoning and scheduling cost for comparable grounding quality. DeepSeek-V4 uses relatively few steps yet performs poorly on both metrics. Video-aware settings also do not necessarily improve the efficiency–quality trade-off.

Search dynamics and controller behavior. We analyze multi-round search through recall trajectories and exploration ratios. As shown in Figure 4, cumulative recall rises sharply in the first three iterations, reaching over 50% of the final recall before marginal gains diminish. This suggests that salient evidence is mostly found early, while later rounds retrieve long-tail or redundant evidence. Models differ in their ability to convert additional iterations into recall gains. Qwen3.5-397B-A17B achieves the highest final recall and continues improving in the middle stage, though with more controller rounds. Gemini3.1-Pro and GPT-5.2 reach strong recall earlier and quickly saturate. In contrast, Qwen3.5-35B-A3B attains lower final recall as its active tasks decline early, while DeepSeek-V4 sharply reduces active tasks mid-search, limiting further improvement.

Figure 5 further shows how controllers balance exploration, i.e., issuing new search keywords, and exploitation, i.e., continuing retrieval with existing keywords. Qwen3.5-397B-A17B performs strong early exploration and then gradually shifts toward exploitation, which is consistent with its higher final recall. Qwen3.5-35B-A3B shifts toward exploitation earlier and obtains lower recall, suggesting insufficient search diversification. Gemini3.1-Pro and GPT-5.2 maintain relatively high explo-

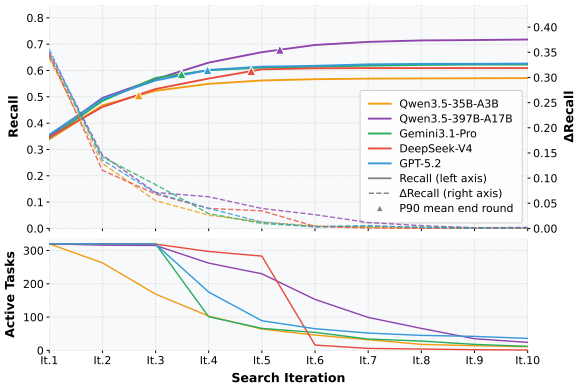

Figure 4: Recall dynamics over search iterations in the Transcript-Only setting. Top: cumulative recall and marginal ΔRecall. Bottom: active search tasks. Triangles mark the mean stopping round, excluding the slowest 10% of tasks.

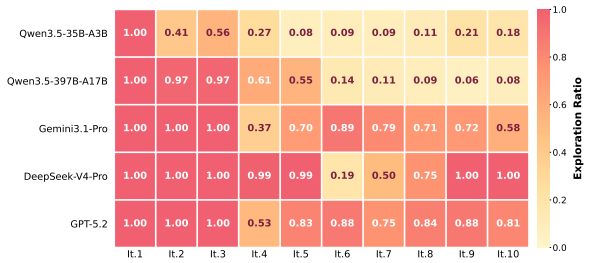

Figure 5: Exploration ratio over search iterations in the Transcript-Only setting. Each cell shows the fraction of active tasks issuing new search keywords in the current iteration. Darker cells indicate stronger exploration and lighter cells indicate more exploitation.

ration in later iterations, but their marginal recall gains are already small, indicating that late exploration is not always cost-effective.

Conditional Benefits of Video Modality. While Video-Aware inputs underperform Transcript-Only inputs on the full benchmark, this does not imply that visual information is unnecessary. We therefore manually curate a diagnostic subset of 30 visually grounded tasks, where at least one major reference viewpoint depends on visual information that cannot be reliably recovered from the transcript. Representative cases include screen-recorded operations such as computer system cleanup, where interface targets and action sequences must be observed on screen, and visually demonstrated procedures such as cooking or styling, where observable states, appearance, or manipulation details are central to the judgment. As shown in Figure 6, Video-Aware inputs improve both Grounded F1 and Content Score for both models,

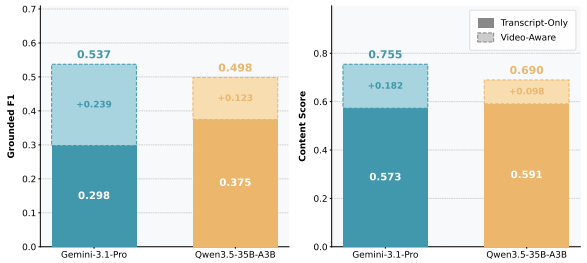

Figure 6: The benefits of video-aware input on visually grounded tasks. Bars compare Transcript-Only and Video-Aware settings on a diagnostic subset of 30 tasks selected from the full benchmark. Dashed segments show the gains from adding video input.

with gains of +0.239/+0.182 for Gemini-3.1-Pro and +0.123/+0.099 for Qwen3.5-35B-A3B. These results indicate that video is a selective evidence source rather than a universally beneficial augmentation: its value becomes evident when answering the query requires visually grounded information.

5 Conclusion

We introduce REELS, a benchmark for evaluating multimodal opinion aggregation in generative search over user-generated videos. It comprises 2,710 real videos across 66 topics and seven domains, with 662 queries covering four opinion-seeking tasks: entity discovery, factor induction, multi-entity comparison, and stance aggregation. Each query includes structured, corpus-grounded reference answers with normalized viewpoints, evidence snippets, source attribution, and source counts, supporting evaluation of both search process and final answer quality. Experiments reveal two key limitations of current models: recall-preserving evidence acquisition and faithful, distribution-aware aggregation. While models often retrieve relevant candidates and understand individual videos, they still struggle to recover opinion distributions across sources. REELS provides a controlled testbed for advancing grounded multimodal search, calibrated agentic retrieval, and distribution-aware opinion aggregation.

Limitations

REELS has several limitations. First, the benchmark is built from Bilibili, a large-scale Chinese user-generated video platform. Although this setting provides rich real-world opinion-seeking scenarios, the current benchmark is limited to Chinese videos and platform-specific discourse patterns. Fu-

ture work can extend the benchmark to multilingual and cross-platform video corpora.

Second, our controlled corpus and simulation retriever improve reproducibility, but they do not fully capture the dynamics of live web search, where videos are continuously added, removed, edited, and re-ranked. This design intentionally prioritizes diagnostic evaluation over real-time deployment realism.

Third, reference construction relies on automatic viewpoint extraction and aggregation followed by human verification. While verification improves reliability, open-ended opinion aggregation remains inherently ambiguous, and some subtle or implicit viewpoints may still be missed or merged imperfectly.

Finally, although REELS is designed for multimodal user-generated videos, not all potentially useful signals are exhaustively modeled. Fine-grained visual cues, tone, editing style, comments, and danmaku may provide additional evidence for user opinions and remain important directions for future work.

Ethical Concerns

Throughout the construction and evaluation of REELS, we treat ethical considerations as an important design principle. Since the benchmark is built from publicly accessible user-generated videos, we avoid collecting private or restricted-access content and focus on aggregated, query-relevant viewpoints rather than profiling individual creators or users. Personally identifying or privacy-sensitive information that is irrelevant to the task is excluded whenever detected. Because user-generated opinions may contain bias, offensive content, or platform-specific preferences, reference answers are manually verified for relevance, evidence faithfulness, and inappropriate content, and unsupported or harmful items are removed when they are not necessary for evaluation. We emphasize that REELS is designed to evaluate corpus-grounded opinion aggregation, meaning that its answers reflect viewpoints expressed in a fixed corpus rather than universal facts or normative judgments. Human annotators are provided with task instructions, training, limited work sessions, and compensation, and may flag sensitive cases for review. Finally, we acknowledge the potential misuse of opinion aggregation systems for large-scale monitoring or profiling, and encourage users of the benchmark to respect platform terms, avoid individual-level surveillance, and report results in a way that preserves the distinction between corpus-specific opinion distributions and broader social conclusions.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reinald Kim Amplayo, Arthur Brazinskas, Yoshi Suhara, Xiaolan Wang, and Bing Liu. 2022. Beyond opinion mining: Summarizing opinions of customer reviews. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3447–3450. ACM.

Stefanos Angelidis and Mirella Lapata. 2018a. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Stefanos Angelidis and Mirella Lapata. 2018b. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3675–3686. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020a. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5151–5169. Association for Computational Linguistics.

Arthur Brazinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5151–5169. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive

summarization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, Proceedings of Machine Learning Research, pages 1223–1232. PMLR.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *CoRR*, abs/2506.11763.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177. ACM.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Bo Pang and Lillian Lee. 2007. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13691–13701. Computer Vision Foundation / IEEE.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

A Appendix

A.1 Overview of Seven Domains

1. **Technology:** This domain covers technology-oriented content related to digital devices, consumer electronics, software tools, artificial intelligence, and automotive technology. It includes reviews, usage experiences, feature

analyses, product comparisons, and discussions of technological trends. Content in this domain often emphasizes practical evaluation, technical functionality, user experience, and the broader implications of emerging technologies.

2. **Fashion:** This domain focuses on personal appearance, style, and beauty-related practices. It includes skincare, makeup, clothing, accessories, fragrance, hairstyling, and beauty routines. Content in this domain often combines experiential advice, aesthetic judgment, product evaluation, and practical tutorials, with an emphasis on individual preferences, cost-effectiveness, and visible outcomes.

3. **Lifestyle:** This domain encompasses everyday consumption, leisure activities, and personal life experiences. It includes food, travel, home improvement, pets, parenting, local services, and recreational experiences. Content in this domain is typically grounded in daily practice and subjective experience, often involving recommendations, warnings, comparisons, and reflections on convenience, enjoyment, value, and quality of life.

4. **Wellness:** This domain covers physical well-being, exercise, sports, fitness, medical and health-related experiences, and lifestyle-based health management. It includes discussions of training methods, body conditioning, sleep, diet, mental well-being, and health-related habits. Content in this domain often combines personal experience, practical advice, and perceived effectiveness, while requiring attention to variation across individuals and contexts.

5. **Knowledge:** This domain focuses on knowledge sharing, practical skills, learning strategies, and explanatory content. It includes arts and crafts, study methods, workplace skills, finance and economics, history, humanities, and general educational knowledge. Content in this domain often aims to explain concepts, teach procedures, summarize experiences, or provide actionable guidance, with an emphasis on clarity, usefulness, and knowledge transfer.

6. **Entertainment:** This domain covers leisure-oriented cultural and entertainment content, including film, television, music, celebrities, video games, esports, animation, and online fandom culture. Content in this domain often involves interpretation, evaluation, recommendation, emotional response, and community-based discussion. The opinions expressed are frequently subjective and culturally situated, reflecting diverse user preferences and audience experiences.

7. **Society:** This domain includes public issues, current events, policy discussions, social phenomena, and widely discussed online topics that are not tied to a specific vertical domain. Content in this domain often involves commentary, interpretation, public sentiment, and competing viewpoints on social or political developments. It requires systems to capture disagreement, contextualize claims, and distinguish between factual reporting, personal opinion, and broader societal interpretation.

A.2 Overview of 67 topics

See Table 4.

B Human Evaluation Details

We provide additional details about the human evaluation and reference validation procedure used in REELS. The goal of human evaluation is to ensure that the automatically constructed reference answers are relevant to the query, faithful to the source videos, complete with respect to the retrieved evidence, and properly merged into normalized viewpoint items.

Annotator Recruitment. We recruited 25 graduate students with backgrounds in natural language processing, information retrieval, multimodal understanding, and data annotation. All annotators were Master's or Ph.D. students with prior experience in at least one of the following areas: information extraction, opinion mining, video understanding, retrieval-augmented generation, or human annotation for NLP datasets. Before participating in the formal evaluation, annotators were required to complete a training and qualification process.

Training and Qualification. The training phase consisted of a 4-hour session covering the task definition, annotation interface, reference answer format, and evaluation guidelines. Annotators were introduced to the four task types in REELS: Entity Discovery and Ranking, Factor Induction and Aggregation, Multi-Entity Comparative Judgment,

Domain	Topics
Lifestyle	
	Beijing Duck Recs, NE China Cuisine, Xinjiang Travel, Arctic Travel Log, Egypt Travel Scams, Japan Work Life, N. Korea Travel, Instant Noodle Tips, Bathroom Reno Guide, Haidilao Guide, Zibo BBQ Status, Shunde Food Tour, Chicken Pot Recipe
Technology	
	Firefly Car Review, 2025 Car Wishlist, M4 iPad Pro Review, QQ Dark History, Vision Pro Review, AI Trends, iPhone 17 Review, China Auto Rise, Xiaomi SU7 Test
Fashion	
	Eyebrow Trimming, Wig Buying Guide, Foundation Pitfalls, Skincare Truths, Skincare Essentials, Color Matching Tips, Maillard Makeup, Perfume Recs, Dr. Martens Styling
Entertainment	
	MC Biome Collection, C-Pop Retrospective, Genshin Beginner, Ne Zha 2 Review, Arknights Story, Shawshank Analysis, LoL Worlds Analysis, Chang'an Film Review, Black Myth Wukong
Knowledge	
	C Drive Cleanup, Acrylic Marker Guide, E. Asian Parenting, China Fiscal System, Handmade Soap, Brain Hack: Hard Tasks, Pangdonglai Analysis, Veggie Cutting Guide, Packing Tips
Wellness	
	Lifestyle Fat Loss, Insomnia Remedies, Arm Workout Plan, Posture Correction, Sugar-Free Drinks, Skiing Guide, Sleep Deprivation, Mental Burnout
Society	
	V-Day Parade Review, Rural Elderly Issues, Food Delivery Wars, Youth Dating Decline, Trump Tariffs, Luckin-Moutai Collab, Returnee Reasons, Modi Election Win, Korea Youth Jobs, Korea Political Crisis

Table 4: Distribution of 67 topics across 7 domains.

and Proposition Stance Aggregation. The training emphasized three core principles: whether an extracted item is relevant to the user query, whether its supporting evidence is faithful to the source video, and whether semantically equivalent items should be merged into the same normalized viewpoint. The session also included hands-on practice with representative examples from different domains and task types. After training, each annotator completed a qualification test containing 20 diverse instances. Only annotators whose judgments reached at least 90% agreement with expert labels were allowed to proceed to the formal evaluation.

Evaluation Procedure. During the formal evaluation phase, annotators used a specialized annotation platform designed for systematic verification of reference answers. For each query, the platform presented the query, candidate reference items, supporting evidence snippets, source video identifiers, and relevant metadata. Annotators were asked to verify four aspects of each reference answer: (1) *relevance*, whether the item directly answers the query; (2) *evidence faithfulness*, whether the viewpoint is supported by the cited source content; (3) *completeness*, whether important viewpoints supported by the evidence are missing; and (4) *merging quality*, whether semantically equivalent items are correctly grouped and distinct viewpoints are not over-merged. When necessary, annotators corrected normalized labels, revised viewpoint summaries, removed unsupported evidence, split over-merged items, merged duplicate items, or added missing reference items.

To maintain annotation quality and reduce fatigue, each evaluation session was limited to two hours. Instances were distributed across a two-week period, and annotators were assigned overlapping subsets to enable agreement analysis. The platform automatically recorded annotation time, completion status, and revision operations, which were used for quality monitoring.

Quality Control and Expert Review. We used both automatic and manual checks to ensure the reliability of human evaluation. The platform first verified annotation completeness, format consistency, and whether all revised items contained the required fields, including normalized label, viewpoint summary, evidence snippet, source identifier, and source count. We then measured inter-

annotator agreement on overlapping instances. Cases with substantial disagreement, defined as Fleiss' $\kappa < 0.6$, were further reviewed by two authors with experience in multimodal information retrieval and opinion aggregation. The expert reviewers resolved disagreements by checking the original source videos or transcripts and by applying the same relevance, faithfulness, completeness, and merging criteria used in annotator training.

Compensation. Annotators were compensated at a base rate of $30 per hour. To encourage careful and consistent annotation, we provided additional performance-based bonuses according to agreement with expert-reviewed judgments and compliance with the annotation protocol.

C Prompts

Topic Extraction Prompt

You are a professional data analyst. Your task is to extract the core entity and event from the metadata of trending Bilibili videos and keep the result as a short phrase.

You may summarize and rewrite the extracted entity and event when necessary, so that the result follows a more standard expression and is more generally applicable.

If the extracted content includes the name of a specific content creator, remove the creator's name and keep only the core entity and event.

Target granularity: similar to "iPhone17 review" or "Avengers: Endgame film review".

The input data contains the following fields:
- **Category**: The video category, such as cats, digital products, or single-player games.
- **Title**: The main title of the video.
- **Recommendation Reason**: The official or algorithm-generated recommendation text, which usually contains the key highlight of the video.
- **Description**: The detailed description of the video.

Examples:
Input: `[He Tongxue] In-depth experience with iPhone 15: Is this the future?`
Output: `iPhone15 in-depth experience`

Input: `Avengers: Endgame, Thanos's tears`
Output: `Avengers: Endgame plot analysis`

Output Format:
- Return a standard JSON object: `{"videoID": "extracted result"}`

If it is difficult to extract a meaningful core entity and event, output `"NULL"` as the extraction result.

Figure 7: Prompt for Topic Extraction

Topic Selection Prompt

You are a senior search engine query analyst. Your task is to select high-value topics from a given batch of "popular topics" that are suitable for constructing realistic user search queries.

Selection Criteria:
1. **KEEP**:
- The topic has a clear and commonly understood meaning.
- When users search for it, they usually have a problem-solving intent, such as looking for guides, tutorials, reviews, or information.
- Examples: "Xinjiang travel guide" as a specific activity, "Huawei Mate60" as a specific product, and "how to cook braised pork" as a specific skill.

2. **DISCARD**:
- The topic is overly broad, ambiguous, mainly used for general entertainment, or expresses emotion only.
- It is difficult to construct a specific and aggregatable search query from it.
- Examples: "random challenge" as too broad, "funny moments" as general entertainment, "feeling emo" as emotional expression, and "juejuezi" as semantically meaningless.

Output Format:
You must return only a valid JSON object and must not include any markdown formatting.
The JSON structure is as follows:

{ "results": [{ "topic": "Original topic name", "decision": "KEEP" or "DISCARD", "reason": "Brief reason", "example_query": "Generated query, or null if the decision is DISCARD" }, ...] }

Figure 8: Prompt for High-Value Topic Selection

Viewpoint Extraction Prompt

Role Assignment
You are a data analyst for Deep Research. Your goal is to extract structured viewpoints from video content for constructing large-scale statistical reports on user opinions.

Task Objective
Extract users' **opinions, evaluations, and experience feedback** about "{task_name}" from the video content, and abstract, summarize, and structure these viewpoints.

Task Focus
- Focus on **opinions, evaluations, and user experience**.
- Do not focus on explicit facts such as searchable parameters, specifications, dimensions, weight, and other factual attributes.

Extraction Rules
1. **Merge similar items**: If multiple detailed descriptions point to the same experience dimension, such as "lagging", "frame drops", or "thermal throttling caused by overheating", merge them into a higher-level dimension such as "performance stability" or "thermal performance".
2. **Aspect must be a reusable general dimension**: The `Aspect` field must be a general evaluation dimension that is coarse-grained and highly summarized, such as price, configuration, intelligent driving, battery life, interior, or after-sales service. Do not use non-reusable detailed terms such as loose screws, loose USB port, or poor third-row armrest. These should be generalized into dimensions such as quality control or space utilization.
3. **The entity must be included**: Both `viewpoint_summary` and `topic` must include the entity being evaluated, and the entity must be written in its full name. Pronouns such as "this car" or "it" are strictly prohibited.
4. **Stay on topic**: All viewpoints must be directly related to "{task_name}".
5. **Filter facts and retain opinions**:
- × Filter out: This car has a wheelbase of 3 meters. This is a parameter that can be easily searched online and does not require video research.
- ✓ Retain: Although this car has a long wheelbase, its rear-seat space utilization is very poor. This is an opinion that requires research.
6. **Evidence Snippet extraction rule**: The evidence snippet must come directly from the original words in the video. Moderate truncation is allowed, but the original wording must not be stylistically polished or rewritten.
7. **Comprehensively extract viewpoints**: Extract as many viewpoints as possible.

Output Format (JSON)
{ "extracted_points": [{ "entity": "Entity", "aspect": "Topic keyword / topic dimension", "viewpoint_summary": "List a series of viewpoints around this topic", "evidence_snippet": "Original quotation excerpt", "topic": "Summarize the topic discussed by this information point" }] }

Example:
... ...
}

Figure 9: Prompt for Structured Viewpoint Extraction from Video Content

User Query Generation Prompt

Role
You are an AI question generation expert who generates high-quality user questions from user viewpoints. Given a set of "user viewpoints", your task is to generate one plausible user question.
Input Format
The input is a set of viewpoint samples extracted from multiple videos. Each sample contains: `aspect`, the viewpoint dimension; `topic`, the discussion topic; and `viewpoint`, the user's evaluation or opinion.
Task Objective
Construct a user question based on these viewpoints. The question must not be answerable by a single factual statement; it should require aggregating multiple user viewpoints. It does not need to cover all viewpoints, as long as it forms a high-quality user query.
Question Type Selection Rules (must be strictly followed)
Select the most suitable task type from the following four categories:
Type A: **Entity Ranking**. Used when the objects are named entities with clear standards and low ambiguity, and the goal is to obtain a Top-K list. Example: "Which budget smartphones around 1,000 RMB are most worth buying in 2025?"
Type B: **Key Factor Extraction**. Used when the object is an abstract concept with non-standardized granularity. Example: "What factors do users care about most when choosing an air purifier?"
Type C: **Multi-Entities Comparative**. Used when multiple entities are explicitly compared, such as Product A vs. Product B, and the viewpoints contain competitive differences or relative evaluations. Example: "Between AITO M9 and Li Auto L9, which one is more recognized by owners for intelligent driving?"
Type D: **Opinion Camp Analysis**. Used when viewpoints discuss a clear judgment, such as whether something is worth buying or whether a configuration is necessary. This type must contain explicit options, such as A/B/C or yes/no, rather than open-ended factor extraction. Example: "Is the rear-seat space of the standard-wheelbase Porsche Panamera sufficient for daily use?"
Question Generation Requirements (very important)
The generated question must satisfy the following requirements:
1. **Task type priority**: Prioritize Entity Ranking if suitable; otherwise use Multi-Entities Comparative if suitable; if neither applies, choose Key Factor Extraction or Opinion Camp Analysis.
2. **Specific entity names must appear**: Do not use pronouns such as "it", "this car", or "this phone".
3. **Multi-viewpoint dependency**: The question must require multiple viewpoint samples to answer; factual questions or questions answerable by a single viewpoint are prohibited.
4. **Reflect shared focus**: The question must reflect the common focus of the samples and must not invent unsupported dimensions or opinions.
5. **Natural user wording**: The question should resemble real questions users ask on video platforms or forums: concise, purposeful, and user-oriented.
6. **Anti-Frankenstein Rule**: If the cluster contains many mixed dimensions, do not combine all of them into one question. Select one core dimension with the most samples, strongest controversy, or clearest focus. Avoid: "How are the screen, charging port durability, and cost-effectiveness of the M4 iPad Pro?" Prefer: "After adopting tandem OLED, what screen display issues have users actually reported for the M4 iPad Pro?"
Output Format (strict JSON)
... ...

Figure 10: Prompt for User Query Generation from Aggregated Viewpoints

D Declaration of LLM Usage

We acknowledge that Cursor was used as an AI-assisted writing tool to polish the language and improve the clarity of an early draft of this manuscript. The tool was used only for grammar correction, wording refinement, and readability improvement. All research ideas, experimental designs, analyses, and conclusions presented in this paper were independently developed, verified, and finalized by the authors.