TCVP: A Practical Pipeline for Video Moment Retrieval Datasets Leveraging Timestamped YouTube Comments

Anonymous ACL submission

Abstract

Video Moment Retrieval (VMR) aims to identify a temporal moment in a video that corresponds to a user query. Existing VMR datasets are typically constructed through either human annotation or machine generation, but both approaches have limitations. Human annotation requires substantial effort for long-form videos, while machine generation produces less salient target moments and queries that resemble captions. To address these limitations, we propose a practical pipeline for VMR dataset generation, named TCVP, where we leverage timestamped YouTube comments to identify interesting moments and capture what viewers focus on. To construct VMR pairs that reflect viewer focus, the pipeline filters uninformative comments (e.g., "07:22 lol") and labels retained comments by the modality of the referenced content. Human evaluation shows that annotators prefer timestamp-selected moments in 98% of cases and comment-guided queries over existing baselines in 92% of cases. Moreover, our modality-aware model evaluation reveals model strengths and limitations across input and retrieval settings. The dataset and code are available in an anonymous repository: https://anonymous.4open.science/r/TCVP.

1 Introduction

Video moment retrieval (VMR) localizes the temporal segment in a video that corresponds to a user query (Lei et al., 2021; Moon et al., 2023; Ren et al., 2024; Pan et al., 2025), enabling users to find desired moments in long videos without extensive manual browsing.

To enable the development and evaluation of VMR models, existing studies (Gao et al., 2017; Lei et al., 2020, 2021; Lin et al., 2023; Soldan et al., 2022) have introduced VMR datasets. These datasets construct moment–query pairs primarily through human annotation or machine generation. Human-annotated datasets ask annotators to watch videos, write queries, and mark the moments that match those queries. This design yields human-curated moment-query pairs, but extending it to many long videos requires substantial annotation effort. Machine-generated datasets reduce this effort by segmenting videos into clips and generating caption-like queries for each clip. However, this scale comes from treating every segmented clip as a retrieval target. As a result, such datasets can include target moments that viewers may not naturally notice.

To address this limitation, we use timestamped YouTube comments to select target moments and guide query generation. A timestamped comment includes a time marker, such as 07:22, that points to a specific moment in the video. These comments indicate which moments viewers found worth mentioning and what they noticed there. We therefore propose the **T**imestamped **C**omment-guided **V**MR **P**ipeline (**TCVP**), which converts these signals into VMR pairs by filtering irrelevant comments, assigning visual-related, audio-related, or mixed labels, and generating queries from the comment focus.

To validate the proposed pipeline, we first assess the reliability of its modality labels by measuring acceptance rates from independent LLM judges and human annotators. We then analyze modality-label distributions across video domains to characterize how viewer focus differs by domain. To assess whether timestamped comments provide practical signals for VMR data construction, we conduct human evaluation comparing timestamp-selected moments and comment-guided queries with baseline methods.

Finally, we benchmark a range of existing VMR models on our dataset, revealing how model strengths and limitations vary across input modalities and retrieval methods.

Our contributions are summarized as follows:

Dataset	Moment Grounding	Query Generation	Modality Label	Dur. (s)	#Queries	Domain
CharadesSTA (Gao et al., 2017)	Human annotation	Human-written	–	30.6	3.7k	Activity
TVR (Lei et al., 2020)	Human annotation	Human-written	–	76.2	5.5k	TV show
QVHighlights (Lei et al., 2021)	Human annotation	Human-written	–	150.0	1.5k	Vlog / News
Ego4D-NLQ (Grauman et al., 2022)	Human annotation	Human-written	–	493.7	4k	Egocentric
MomentSeeker (Yuan et al., 2025)	Human annotation	Human-written	–	1201.9	1.8k	Open
LongVALE (Geng et al., 2025)	Auto segmented	Auto-generated	–	235.0	105k	Open
TriSense-2M (Li et al., 2026)	Auto segmented	Auto-generated	–	905.0	2M	Open
Ours	Viewer timestamp	Auto-generated	Visual / Audio / Mixed	1503.5	17.1k	YouTube

Table 1: **Overview of VMR datasets.** We compare existing VMR datasets from the perspective of dataset design and practical utility. Moment grounding indicates how target moments are selected, either by annotators, automatic segmentation, or viewer timestamps. Query generation indicates whether queries are human-written or automatically generated, and modality label indicates whether explicit query-level modality labels are provided.

- We introduce a novel perspective on VMR dataset construction by leveraging timestamped YouTube comments.
- We construct a long-form YouTube VMR dataset with modality labels and analyze domain-specific viewer focus.
- We show that TCVP constructs high-quality VMR data from timestamped comments, with human annotators preferring its selected target moments and generated queries.
- We benchmark existing models on our dataset, showing that separating queries by modality label reveals meaningful differences.

2 Existing VMR Dataset Construction

Existing VMR datasets can be broadly grouped into human-annotated and machine-generated approaches. Human-annotated datasets (Gao et al., 2017; Lei et al., 2020, 2021; Yuan et al., 2025) rely on annotators to define both target moments and their queries. Specifically, QVHighlights (Lei et al., 2021) is constructed by asking annotators to write a query for each video and mark the matching time spans and rate how important each clip is for that query. This procedure lets annotators define the retrieval targets directly, but scaling it to many long videos requires substantial annotation effort. For long-form content, annotators may also have limited context for judging which moments stand out across the full video.

Machine-generated datasets take a different route by dividing videos at visual or audio change points and generating query-like descriptions for every resulting clip (Geng et al., 2025; Li et al., 2026). This procedure reduces annotation effort and can produce moment–query pairs at scale. However, treating every clip as a target moment can include clips that are easy to segment but not moments that viewers would choose to mention. The generated descriptions can therefore describe clip content without indicating whether the clip is meaningful as a retrieval target.

Table 1 reflects this trade-off. Human-annotated datasets tend to be smaller because annotators must choose moments and write queries. Machine-generated datasets scale more easily by dividing videos into clips and generating query-like descriptions for each clip. However, this scale comes from treating every segmented clip as a target moment. As a result, machine-generated datasets can include clips that are easy to segment but not necessarily moments that viewers would choose to mention.

TCVP lies between these two construction paradigms. By using timestamped comments written by viewers who have already watched the video, it avoids fully manual moment and query annotation while also avoiding the assumption that every segmented clip is a retrieval target. The resulting dataset is more scalable than fully human-annotated datasets, but smaller than machine-generated benchmarks. In return, its target moments are selected using explicit timestamp signals provided by viewers. Broader related work on VMR models is discussed in Appendix A.

3 TCVP: Timestamped Comment-guided VMR Pipeline

We use timestamped YouTube comments to select target moments and guide query generation. A timestamped comment can indicate what viewers focus on at a specific moment in a video. However, not every comment can be used directly for query generation. Some comments are generic reactions that do not refer to the video content, such as "lol" or "this is amazing." Other comments mention the moment, but they may focus on visual content, audio content, or both. This distinction changes what information should be included in the query. For example, in a podcast video where the same

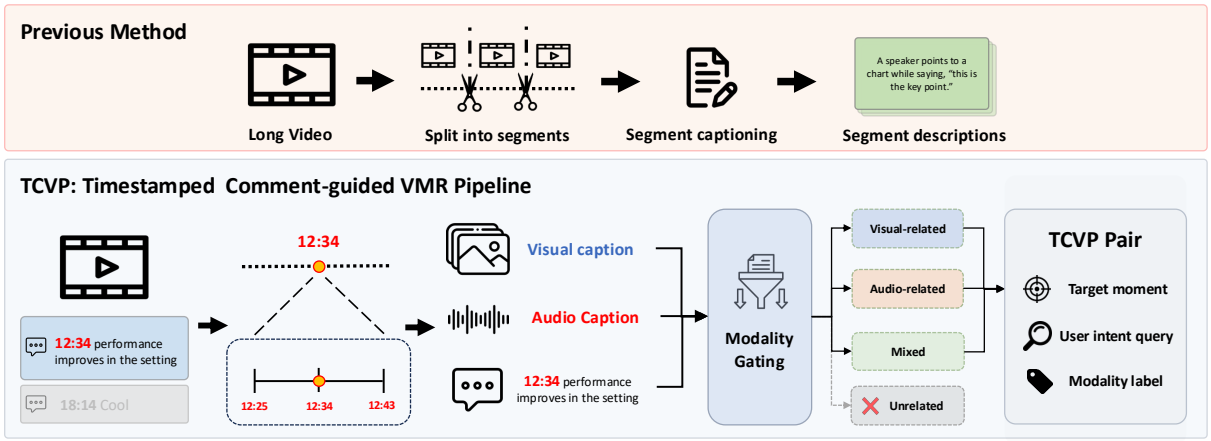

Figure 1: **Overview of existing and proposed VMR dataset generation approaches.** The upper panel illustrates a typical segment-based approach, in which a long video is divided into predefined temporal segments and each segment is annotated with descriptive sentences. In contrast, the lower panel presents TCVP, which leverages timestamped comments to identify candidate moments and to guide query generation. Visual and audio captions extracted around each timestamp are used to filter irrelevant comments and assign modality labels—visual, audio, or mixed—thereby yielding a VMR instance consisting of a target moment, a user-intent query, and a modality label.

speaker remains on screen for a long time, a visual description alone may not distinguish the target moment. If a comment focuses on a quoted line, the query should preserve that audio-related focus instead of describing the speaker's appearance. To address these issues, we propose the **T**imestamped **C**omment-guided **V**MR **P**ipeline (**TCVP**). TCVP filters comments that do not refer to the video content, uses modality gating to determine whether each retained comment focuses on visual content, audio content, or both, and generates a query from the retained comment focus. The pipeline proceeds through four stages: (1) video and comment collection, (2) modality-specific captioning, (3) comment filtering and modality gating, and (4) query generation. We describe each stage below.

Video and comment collection We collect videos from YouTube channels with more than one million subscribers, focusing on long-form videos with an average duration of over 25 minutes with active comment sections. The collected videos cover 20 categories, each detailed with representative channels and descriptions provided in Appendix E. For each video, we collect comments and retain those containing at least one timestamp. When a comment contains multiple timestamps, we use the first timestamp. We resolve duplicate timestamps by selecting the most-liked comment and keep the top 20 timestamped comments per video. This process yields 19,445 timestamped comments from 1,014 videos. Videos without timestamped comments are discussed in Appendix F.2.

Modality-specific captioning Each retained timestamp is treated as a local candidate moment. For each timestamp, we use open-source MLLMs to generate two captions for a symmetric window of half-width δ around it, one describing its visual content and the other its audio content. Keeping these captions separate lets the filtering and modality-gating step determine whether the comment focuses on visual content, audio content, or both. We analyze the effect of this local captioning window in Appendix F.

Comment filtering and modality gating In this step, TCVP labels each timestamped comment as visual-related, audio-related, mixed, or unrelated. The unrelated label is used to filter out comments that are not suitable for retrieval query generation. The label is determined based on what moment the comment refers to and whether that moment can be identified from the visual or audio captions. Generic reactions such as "lol" or "this is amazing," comments unrelated to the video content, and comments whose target cannot be identified from the captions are labeled as unrelated and removed. For the remaining comments, the modality label reflects the modality that is central to the commented moment. For example, in a podcast-style clip where the visual scene changes little and the spoken content carries the relevant information, a comment that points to the phrase "I hate to tell you" is labeled as audio-related. Table 2 shows representative outputs of this step. TCVP keeps the part of each comment that is supported by the visual or au-

Timestamped comment	Label
6:55 you can see it on her face *11:33 that backwards shot goes in without touching the hoop*	Visual
3:19 Lee calling Alex "Tree Bark" is an underrated line *29:40 the way he says "I hate to tell you" made me laugh so hard*	Audio
2:01 the text says INDENTIFY right when Ben says "identify" *2:53 Father's expression while Peele says he cared for his parents for 15 years is priceless*	Mixed
4:13 thanks *11:13 This is amazing*	Unrel.

Table 2: **Examples of timestamped comments and modality labels.** Selected examples illustrate how timestamped user comments capture visual, audio, or mixed information. Visual and audio cues are highlighted in red and blue, respectively.

dio captions, assigns a modality label, and removes unrelated comments before query generation.

The resulting modality label determines which content is used in the query generation stage and is stored as the label of the final moment–query pair. After this step, TCVP retains 17,147 comments from 19,445 timestamped comments and removes 2,298 uninformative comments. The retained comments and their modality labels are then used for retrieval query generation.

User-intent query generation We generate a final query from the comment, its modality label, and the corresponding video captions. Because the query is guided by a viewer-written comment, it reflects the viewer's intended focus rather than an artificially constructed description (Geng et al., 2025; Li et al., 2026). The modality label determines whether the query is grounded in visual evidence, audio evidence, or both. The prompts used in our pipeline are provided in Appendix G.

Implementation Details For window captioning, we use Qwen2.5-Omni (Xu et al., 2025) and set $\delta = 9$s, yielding an 18-second window around each timestamp. We analyze the effect of our local window design in Section F.1. For query generation, we prompt the proprietary LLM `GPT-5 mini` (OpenAI, 2026) with a structured prompt.

4 Analysis of TCVP

This section analyzes the moment–query pairs produced by TCVP from three perspectives. The first analysis audits whether the filtering results and modality labels are reliable. The second examines how modality labels are distributed across video domains. The third measures human preferences for the selected moments and generated queries.

Judge	Avg.	Vis.	Aud.	Mix.	Unrel.
GPT-5.4	95%	100%	92%	92%	96%
Claude Opus 4.6	96%	100%	84%	100%	100%
Gemini 3.1 Pro	95%	96%	88%	96%	100%
Human	98%	96%	96%	100%	100%

Table 3: **Acceptance rates for modality label.** We report whether independent evaluators judged the TCVP-assigned label as appropriate for the timestamped comment and local video evidence.

4.1 Filtering and modality label reliability

Our pipeline assigns each timestamped comment one of four labels: visual-related, audio-related, mixed, or unrelated. Since these labels are produced by an LLM-based pipeline, we examine their reliability with independent evaluators.

We construct a validation subset by sampling instances evenly across the modality labels. We evaluate this subset using three independent LLM judges, GPT-5.4 (OpenAI, 2026), Claude Opus 4.6 (Anthropic, 2026), and Gemini 3.1 Pro (Google DeepMind, 2026), together with three human annotators. Each evaluator receives the timestamped comment, the modality label assigned by TCVP, and information about the video content around the timestamp. They make a binary judgment on whether the assigned label is appropriate.

Table 3 shows high acceptance across both LLM judges and human annotators. The three LLM judges accept 95–96% of the labels on average, indicating that the TCVP labels are largely judged appropriate by independent models. Human annotators show an overall acceptance rate of 98%, with 96% acceptance for both visual-related and audio-related labels and 100% for mixed labels. The unrelated labels receive 96–100% acceptance from all LLM judges and 100% acceptance from human annotators. This indicates that comments labeled as unrelated are consistently judged as not being grounded in the local video content, supporting the reliability of the filtering step. Overall, these results indicate that TCVP's labeling decisions are largely consistent with independent evaluator judgments. Detailed sampling procedures, subset composition, prompts, and annotation protocols are provided in Appendix H.1.

4.2 Modality Label Distribution

Following the assessment of modality label reliability, we examine how these labels are distributed

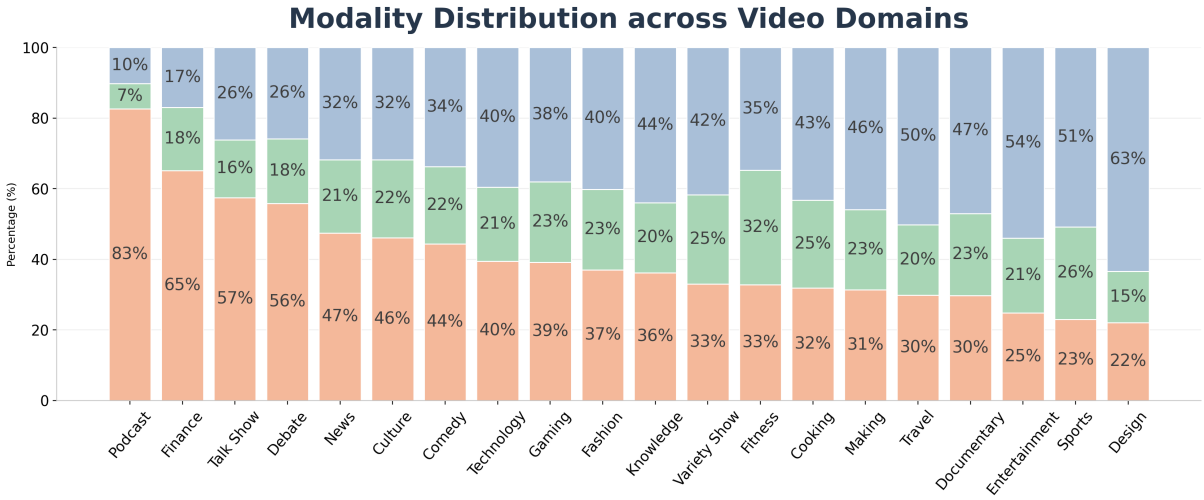

Figure 2: **Modality distribution across video domains.** We collect videos from diverse domains and report the proportion of modality labels within each domain. The stacked bars show label proportions by domain, with orange, green, and blue denoting audio, mixed, and visual labels, respectively.

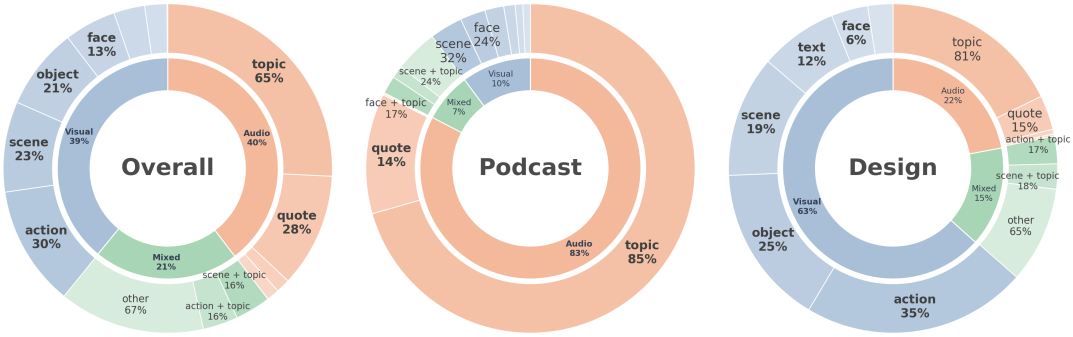

Figure 3: **Modality and focus-type distributions.** We report distributions for the overall retained set and for Podcast and Design, the domains with the strongest audio and visual skews. The inner ring shows modality-label proportions, while the outer ring decomposes each label into finer-grained focus types.

across video domains such as Podcast, Design, and Sports (Figure 2). We then analyze finer-grained focus types within each label to characterize what viewers attend to in each modality (Figure 3).

Across the entire dataset, visual-related and audio-related pairs in nearly equal proportions, accounting for 39.1% and 39.6% of the total instances, respectively, while mixed pairs make up the remaining 21.3%. However, Figure 2 shows that this distribution varies substantially across video domains. Audio-centered domains, such as Podcast, Finance, and Talk Show, contain a much larger proportion of audio labels. In contrast, visually oriented domains, such as Design and Sports, contain more visual labels.

Viewer focus can also differ within the same modality label. Figure 3 illustrates this by presenting finer-grained focus types in the domains most strongly skewed toward audio and visual labels. In Podcast, where 82.6% of retained pairs are labeled as audio, most audio-focused comments refer to what speakers discuss, with spoken topics accounting for 85% and direct quotes accounting for 14%. This indicates that viewers of speech-centered content primarily attend to the substance of what is said. In contrast, Design shows a more diverse visual focus. Within its visual labels, viewer focus is spread more evenly across actions, objects, scenes, on-screen text, and faces. These patterns suggest that viewer focus varies not only across video domains, but also in the fine-grained cues associated with each modality.

4.3 Practicality of comment-guided moments and queries

Case studies. Figure 4 shows a moment selection example using a timestamped comment from a horror game play video. The timestamp-selected clip captures a sudden monster attack that leads to a game over, whereas the comparison clip shows the player checking CCTV cameras.

Figure 5 compares generated queries for an ex-

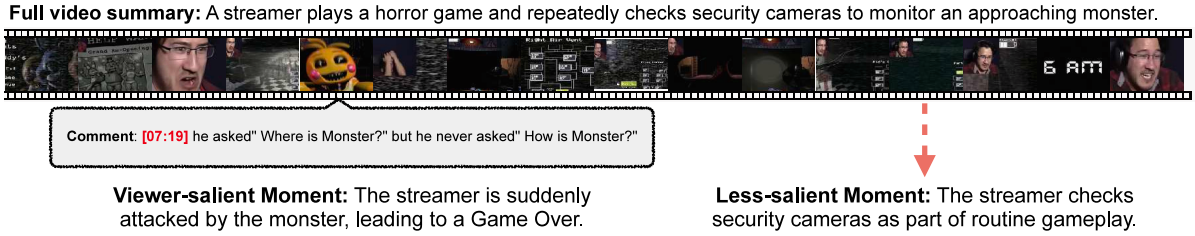

Figure 4: **Effectiveness of timestamp-guided moment selection.** Examples from the same source videos show that timestamped comments provide viewer-grounded cues for localizing salient events. The comparison clips are sampled from non-overlapping timestamps in the same video and often capture more routine content.

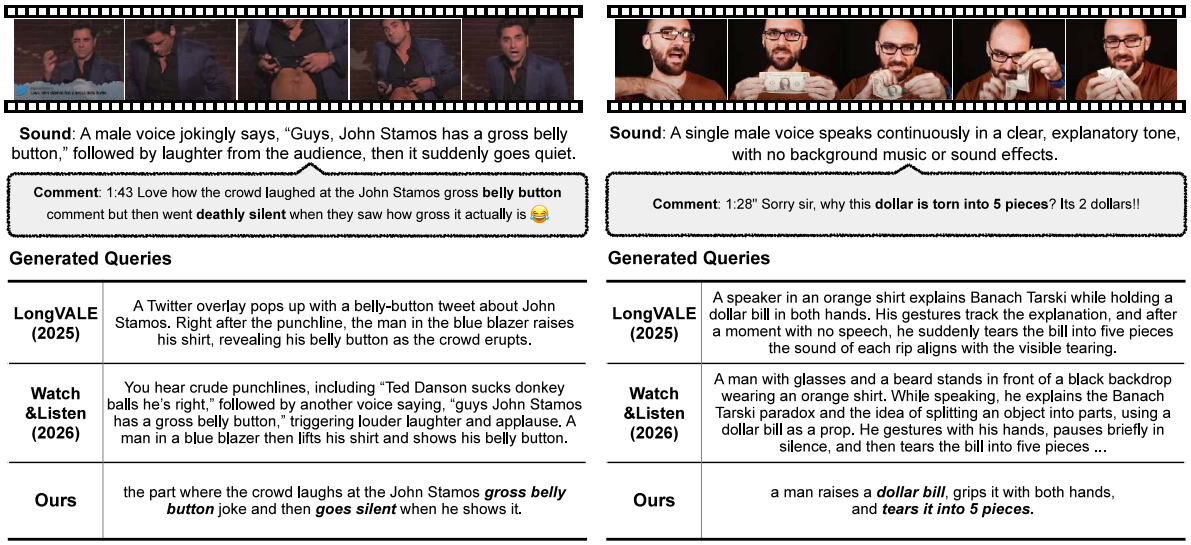

Figure 5: **Examples of user-intent query generation.** We compare generated queries from existing methods and our approach. Existing methods tend to produce descriptive queries that closely resemble visual or audio captions, whereas our approach leverages comments to better reflect user intent, resulting in more natural queries.

Evaluation	Preferred option	Preference
Moment selection	Timestamp selected	98%
	Random selected	2%
Query generation	Ours	92%
	LongVALE	5%
	Watch&Listen	3%

Table 4: **Human preference evaluation.** Moment selection compares a timestamp-selected clip with a non-overlapping random clip from the same video. Query generation compares queries for the same target moment. Percentages indicate majority-vote human preferences.

ample where the timestamped comment refers to a dollar bill being torn into five pieces. LongVALE and Watch&Listen generate caption-like queries that describe the surrounding explanation and the dollar bill. Ours reflects the user's focus in the comment and generates a user-intent query about the dollar bill being torn into five pieces.

Human evaluation. We evaluate 100 videos with three annotators, aggregate preferences by majority vote, and report the results in Table 4. Appendix H provides the annotation protocol.

For moment selection, annotators compare two clips from the same video. One clip is centered on a timestamped comment, while the other is sampled from a non-overlapping random moment in the same video. This comparison tests whether timestamped comments identify viewer-salient moments more effectively than random alternatives from the same video. Annotators prefer the timestamp-selected clip in 98% of videos.

For query generation, annotators compare queries generated for the same target moment by LongVALE (Geng et al., 2025), Watch&Listen (Li et al., 2026), and Ours. They choose the query they would most likely use to find the moment again. Annotators prefer the queries generated by TCVP in 92% of cases.

These results show that human annotators prefer comment-guided moments and generated queries, supporting the practicality of timestamped comments for VMR dataset construction.

Model	Size	Input			Visual-related			Audio-related			Mixed		
		V	A	S	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Direct Timestamp Prediction													
Gemini 2.5 Flash (Comanici et al., 2025)	–	✓			6.0	–	–	4.0	–	–	8.0	–	–
	–	✓	✓		6.0	–	–	6.0	–	–	8.0	–	–
Qwen2.5-Omni (Xu et al., 2025)	3B	✓	✓		5.2	–	–	3.7	–	–	6.4	–	–
Qwen2.5-VL (Bai et al., 2025)	3B	✓			6.1	–	–	4.6	–	–	7.2	–	–
InternVL3 (Zhu et al., 2025)	2B	✓			3.7	–	–	2.5	–	–	4.8	–	–
Segment-caption Retrieval													
Qwen2.5-Omni (Xu et al., 2025)	3B	✓	✓		15.9	35.9	50.6	10.5	24.4	34.4	21.0	42.9	54.4
	3B	✓	✓	✓	15.4	38.3	51.7	**30.3**	**56.2**	**67.1**	28.8	55.8	66.4
Qwen2.5-VL (Bai et al., 2025)	3B	✓			10.8	30.8	42.6	4.6	14.4	23.4	10.6	27.9	39.1
	3B	✓		✓	12.0	33.4	46.3	29.7	54.6	63.9	24.6	50.2	63.3
InternVL3 (Zhu et al., 2025)	2B	✓			11.1	27.7	39.3	4.5	13.9	22.9	11.2	27.7	37.0
	2B	✓		✓	12.2	30.7	42.6	23.1	42.3	52.7	21.5	43.1	52.9
Embedding Retrieval													
CLIP (Radford et al., 2021)	151M	✓			20.4	50.2	65.2	7.9	22.6	33.9	19.3	47.6	61.1
LanguageBind (Zhu et al., 2024)	428M	✓			**28.7**	**62.2**	**74.9**	12.2	30.1	40.8	**28.9**	**60.9**	**73.1**
InternVideo2 (Wang et al., 2024)	1B	✓			23.9	54.7	68.4	7.0	21.6	33.1	21.5	50.4	64.4
CLAP (Elizalde et al., 2023)	154M		✓		3.6	15.2	26.7	3.7	13.9	24.2	5.9	20.1	32.2
LanguageBind-Audio (Zhu et al., 2024)	428M		✓		3.3	15.3	27.7	3.1	13.8	24.1	5.1	19.4	31.0

Table 5: **Retrieval performance on the dataset generated by TCVP.** We evaluate existing VMR models across visual-related, audio-related, and mixed queries. The results show that multimodal inputs generally improve retrieval performance, while audio-related queries remain particularly challenging. V, A, and S denote visual frames, raw audio, and subtitles, respectively.

5 Evaluation of Models on Our Dataset

We evaluate existing models on TCVP for target moment retrieval from user-intent queries.

Task and metrics. Each evaluation item pairs a user-intent query with a target timestamp t_i in a video. The model is required to retrieve the video moment corresponding to the query. We report results separately by modality label. For direct timestamp prediction, the model outputs a single timestamp $\hat{t}_i$. A prediction is considered correct if $|\hat{t}_i - t_i| \leq 10$ seconds, and performance is reported using Recall@1. For segment-based and embedding retrieval, each video is divided into non-overlapping 10-second segments. Recall@K measures whether the top-K retrieved segments include the segment containing t_i.

Evaluation settings. Table 5 compares three evaluation settings. In direct timestamp prediction, multimodal language models (MLLMs) directly predict a timestamp from the query and the provided video inputs. For visual input, all models in the same setting receive the same 100 frames uniformly sampled from the full video. We evaluate the closed-source MLLM Gemini 2.5 Flash (Comanici et al., 2025) and three open-source MLLMs: Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and Qwen2.5-Omni (Xu et al., 2025).

In segment-caption retrieval, we use the same open-source MLLMs to convert each 10-second segment into a text description before retrieval. When subtitles are used, all models use the same YouTube-provided subtitle text aligned to each 10-second segment. The query and segment descriptions are then embedded with Qwen-3 Embedding (Zhang et al., 2025), and segments are ranked by cosine similarity.

The embedding retrieval setting encodes each 10-second segment into either a visual or audio embedding and compares it with the query. We use CLIP (Radford et al., 2021), LanguageBind (Zhu et al., 2024), and InternVideo2 (Wang et al., 2024) for visual embeddings, and CLAP (Elizalde et al., 2023) and LanguageBind-Audio (Zhu et al., 2024) for audio embeddings.

5.1 Main Results

Direct timestamp prediction. Both the closed-source model and the open-source MLLMs show low recall in direct timestamp prediction. With visual input alone, all models remain below 8.0 R@1 across modality labels. Adding raw audio to Gemini 2.5 Flash improves audio-related queries only slightly, from 4.0 to 6.0 R@1, while leaving the other subsets unchanged. These results indicate that frame sampling from a long video is difficult for direct timestamp prediction.

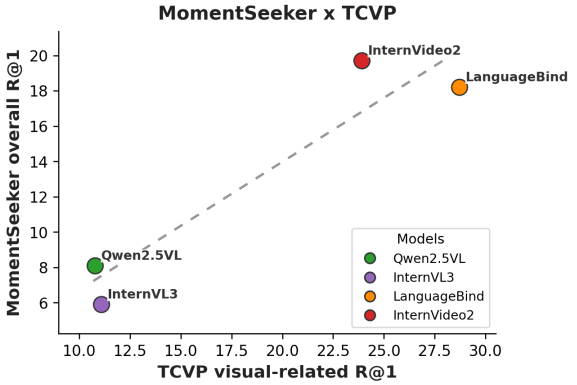

Figure 6: **Consistency with an existing VMR benchmark.** Model performance on TCVP is compared with MomentSeeker (Yuan et al., 2025) across four shared models. The positive trend suggests consistency with existing VMR evaluation.

Segment-caption retrieval. Among segment-caption retrieval methods without subtitles, Qwen2.5-Omni performs best overall. This is consistent with its use of both visual frames and raw audio when generating segment descriptions. When subtitles are appended to the segment descriptions, performance improves most clearly on audio-related queries. For Qwen2.5-Omni, audio-related R@10 increases from 34.4 to 67.1, with similar gains for Qwen2.5-VL and InternVL3. These gains indicate that YouTube-provided subtitle text is useful for audio-related retrieval. This matches the analysis in Section 4.2, where many audio-related queries focus on spoken topics or quoted lines. With subtitles, Qwen2.5-Omni also performs strongly on mixed queries, reaching 66.4 R@10.

Embedding Retrieval. Visual embedding models perform strongly on visual-related and mixed queries. LanguageBind achieves the best results among embedding models, reaching 74.9 R@10 on visual-related queries and 73.1 R@10 on mixed queries. InternVideo2 and CLIP show the same pattern, with stronger performance on visual-related and mixed queries than on audio-related queries. Audio-only embedding models show lower performance on audio-related queries. CLAP and LanguageBind-Audio reach 24.2 and 24.1 R@10, respectively, which remains below the segment-caption retrieval setting with subtitles. These results suggest that visual embeddings are effective for visual-related and mixed queries, whereas acoustic embeddings alone may be insufficient for queries centered on spoken topics or quoted lines.

Among the strongest settings on mixed queries, LanguageBind and Qwen2.5-Omni with subtitles show similar performance but different strengths. LanguageBind is also strong on visual-related queries, while Qwen2.5-Omni with subtitles is strongest on audio-related queries. These differences are revealed when results are reported by modality label.

5.2 Comparison with MomentSeeker

We compare TCVP with MomentSeeker (Yuan et al., 2025), an established benchmark focused on long-video retrieval, to examine whether they follow a similar performance trend. For the four models evaluated in both settings, we compare MomentSeeker R@1 with the corresponding TCVP R@1. Figure 6 shows a positive trend between the two benchmarks. As MomentSeeker scores increase, the corresponding TCVP scores also tend to increase.

6 Conclusion

We proposed TCVP, a timestamped-comment-guided pipeline for constructing VMR data from long-form YouTube videos. TCVP uses viewer-provided timestamp signals to select target moments, filters uninformative comments, assigns visual-related, audio-related, or mixed labels, and generates queries that preserve the retained comment focus. Human evaluation shows that TCVP produces high-quality VMR pairs, with annotators preferring timestamp-guided moments in 98% of videos and TCVP queries over baselines in 92% of cases. The resulting modality labels reveal domain-dependent viewer focus and support modality-aware model analysis. Our experiments show that existing VMR approaches differ across input modalities and retrieval methods, with audio-related retrieval benefiting from subtitles and visual embeddings remaining strong for visual-related and mixed queries.

Limitations

TCVP is designed around timestamped comments, so the current dataset focuses on long-form YouTube videos where such comments are available. Extending the same idea to other platforms would require collecting similar moment-level signals, such as retention peaks or platform-generated key moments. Because timestamped comments are anchored to specific moments, they may not fully represent retrieval intents that depend on broader context. We discuss future work in Appendix B.

References

Anthropic. 2026. Claude Opus 4.6. Anthropic Documentation.

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

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

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

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. 2025. Rewind: Understanding long videos with instructed learnable memory. In *CVPR*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, and Renrui Zhang. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2025. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *CVPR*.

Google DeepMind. 2026. Gemini 3.1 Pro model card. Google DeepMind Model Card.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, and Xingyu Liu. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. 2025. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *AAAI*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *ICCV*.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, Wk Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. 2023. Cone: An efficient coarse-to-fine alignment framework for long video temporal grounding. In *ACL*.

Boseung Jeong, Jicheol Park, Sungyeon Kim, and Suha Kwak. 2025. Learning audio-guided video representation with gated attention for video-text retrieval. In *CVPR*.

Junho Kim, Hyunjun Kim, Hosu Lee, and Yong Man Ro. 2025. Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis. In *CVPR*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*.

Pilhyeon Lee and Hyeran Byun. 2024. Bam-detr: Boundary-aligned moment detection transformer for temporal sentence grounding in videos. In *ECCV*.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *ECCV*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*.

Zinuo Li, Xian Zhang, Yongxin Guo, Mohammed Bennamoun, Farid Boussaid, Girish Dwivedi, Luqi Gong, and Qiuhong Ke. 2026. Watch and listen: Understanding audio-visual-speech moments with multimodal llm. In *NeurIPS*.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. 2023. Univtg: Towards unified video-language temporal grounding. In *ICCV*.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2025. Videorag: Visually-aligned retrieval-augmented long video comprehension. In *NeurIPS*.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *CVPR*.

OpenAI. 2026. GPT-5 mini. OpenAI API Documentation. Accessed: 2026-05-25.

OpenAI. 2026. GPT-5.4 model. OpenAI API Documentation.

Junwen Pan, Rui Zhang, Xin Wan, Yuan Zhang, Ming Lu, and Qi She. 2025. Timesearch: Hierarchical video search with spotlight and reflection for human-like long video understanding. *arXiv preprint arXiv:2504.01407*.

Qing Ping and Chaomei Chen. 2017. Video highlights detection and summarization with lag-calibration based on concept-emotion mapping of crowdsourced time-sync comments. In *NFiS@EMNLP*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. 2025. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *ICML*.

Mattia Soldan, Alejandro Pardo, Juan Leon Alcazar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *CVPR*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025a. Lvbench: An extreme long video understanding benchmark. In *ICCV*.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, and Yansong Shi. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *CVPR*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, and Kai Dang. 2025. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.

Linli Xu and Chao Zhang. 2017. Bridging video content and comments: Synchronized video description with temporal summarization of crowdsourced time-sync comments. In *AAAI*.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. 2025. T*: Re-thinking temporal search for long-form video understanding. In *CVPR*.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Zhicheng Dou, and Ji-Rong Wen. 2025. Momentseeker: A task-oriented benchmark for long-video moment retrieval. *CVPR*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, and Junyang Lin. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Junjie Zhou, Yan Shu, Bo Zhao, and Boya Wu. 2025. Mlvu: Benchmarking multi-task long video understanding. In *CVPR*.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, and Zongwei Li. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *ICLR*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, and Jie Shao. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Related Work

Video moment retrieval models. Video moment retrieval aims to locate the time span in an untrimmed video that matches a natural-language query (Gao et al., 2017; Lei et al., 2021). Many recent methods formulate this task as DETR-style span detection, predicting a small set of candidate moments end-to-end and improving query conditioning or boundary alignment (Lei et al., 2021; Moon et al., 2023; Lee and Byun, 2024). Multimodal large language models have also been adapted to temporal grounding through timestamp prediction and timestamp-aware reasoning (Ren et al., 2024; Guo et al., 2025; Liu et al., 2025). For long videos, coarse-to-fine methods first retrieve query-related regions and then refine temporal boundaries (Hou et al., 2023; Pan et al., 2025). Overall, these methods study how to retrieve target moments once a query is given.

VMR dataset construction. Existing VMR datasets construct moment–query pairs through human annotation or automated procedures. Many human-annotated datasets ask annotators to write descriptions for temporal windows or to mark temporal windows for given descriptions (Regneri et al., 2013; Gao et al., 2017; Hendricks et al., 2017; Krishna et al., 2017; Lei et al., 2020, 2021). Related long-video retrieval benchmarks broaden this setting to egocentric episodic-memory queries or task-oriented queries with text, image, or video inputs (Grauman et al., 2022; Yuan et al., 2025). Automated datasets reduce manual effort by aligning existing descriptions with videos or by generating query-like text from local visual, speech, or audio content (Soldan et al., 2022; Geng et al., 2025; Li et al., 2026). Beyond VMR, time-sync comments have been used for highlight detection, summarization, and temporal video description (Ping and Chen, 2017; Xu and Zhang, 2017). TCVP differs by constructing instance-level VMR pairs: timestamps provide candidate retrieval targets, while comments are filtered and grounded to generate searchable queries and modality labels. This distinction matters because viewer saliency does not necessarily imply retrievability; generic or unsupported reactions may mark interesting moments but cannot reliably serve as retrieval queries.

Long video understanding. Long-video understanding benchmarks test whether models can retrieve and reason over information spread across long videos (Fu et al., 2025; Zhou et al., 2025; Wu et al., 2024; Wang et al., 2025a; Yuan et al., 2025). Because long videos contain many frames, clips, and events, recent systems often reduce the cost of finding relevant information. Some methods reduce visual-token redundancy or transfer long-context language modeling ability to video inputs (Li et al., 2024; Zhang et al., 2024; Shen et al., 2025). Others route queries to relevant temporal units, build hierarchical video representations, or frame long-video reasoning as temporal search (Kim et al., 2025; Wang et al., 2025b; Pan et al., 2025; Ye et al., 2025). Retrieval-augmented and memory-based methods store auxiliary text, sparse memories, or learnable memory states to preserve information beyond the limited context of a video-language model (Song et al., 2024; Luo et al., 2025; Diko et al., 2025). Together, these benchmarks and systems make access to relevant moments central to long-video understanding.

Audio-visual alignment and modality use. Several video-language models incorporate audio to understand video moments that cannot be explained by visual frames alone. Audio-visual LLMs encode visual and audio streams and pass both forms of information to a language model, while long-video benchmarks such as LongVALE include visual, audio, language, and event-level signals (Cheng et al., 2024; Geng et al., 2025). Recent work also makes modality integration more selective through query-based connectors or gated audio-guided representations (Li et al., 2026; Jeong et al., 2025). These methods study how models should combine modalities. TCVP addresses a complementary dataset question by labeling whether each retained comment focuses on visual content, audio content, or both. This makes it possible to evaluate when visual frames, speech text, or audio inputs help VMR models.

B Future Work

Future work can extend TCVP to videos without timestamped comments by using other moment-level signals, such as viewer-retention peaks, creator-provided chapters, or platform-generated key moments. These signals offer different trade-offs: retention peaks reflect aggregate viewer interest, while chapters and key moments provide navigational structure but are often coarser and may not contain viewer-written intent. It can also map comment timestamps to content-aware tem-

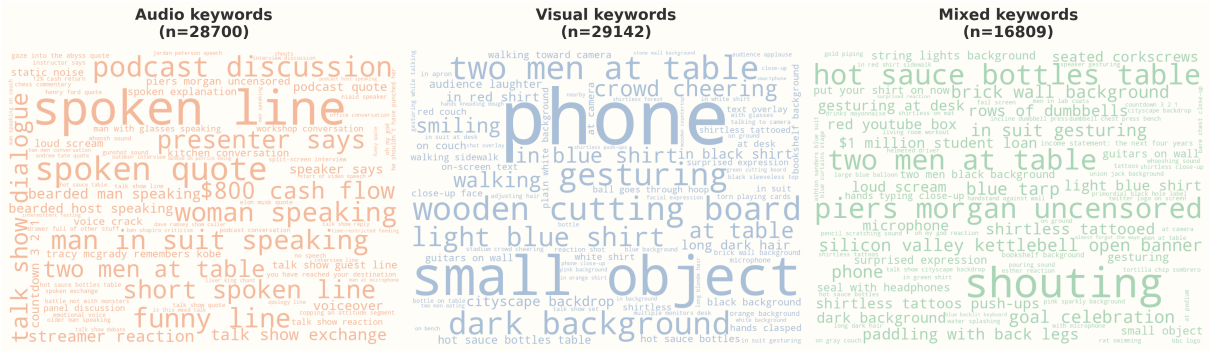

Figure 7: **Modality-specific keyword patterns in retained comments.** We visualize keyword clouds for retained comments grouped by modality label. The clouds show that visual-related, audio-related, and mixed comments emphasize different terms, reflecting distinct types of viewer focus.

poral segments based on visual changes, audio changes, ASR topic shifts, or event boundaries. TCVP pairs can be used to train VMR models, including models that adapt their use of visual frames, speech text, and audio inputs to visual-related, audio-related, and mixed queries. Future work can also use modality-specific failure cases to refine dataset sampling and balance the difficulty of the three query types.

C Ethics Statement

TCVP is constructed from publicly available YouTube videos and timestamped comments. Because timestamped comments are user-generated text, we treat them as data that may contain privacy-sensitive or offensive content even when they are publicly visible. We therefore minimize the amount of user-generated information included in the released artifact. We do not redistribute raw videos, raw comments, commenter IDs, profile information, or raw comment metadata. The released dataset contains only the information needed for research use, such as video identifiers, timestamps, generated queries, modality labels, and processing scripts.

Data collection and release are designed to respect YouTube's Terms of Service and applicable platform policies. Users of the released artifact are expected to access the original videos through the platform and comply with the corresponding terms and policies. Because videos or comments may be edited, removed, or made unavailable by their owners or by the platform, some released identifiers may no longer resolve at a later time.

The dataset is intended for research on video moment retrieval. It should not be used to identify, profile, contact, or infer private attributes of individual users. It should also not be used for moderation, surveillance, or user-level behavioral analysis. The modality labels and generated queries describe retrieval targets and should not be interpreted as judgments about the users who wrote the original comments.

For human evaluation, we recruited annotators through institutional channels. Before annotation, participants received written instructions and were informed that their judgments would be used for research evaluation, reported only in aggregate, and not linked to personal identifiers. Participation was voluntary, and annotators could stop the task at any time. We did not collect demographic information beyond what was needed for compensation. Annotators were compensated based on the estimated task duration, with an estimated hourly rate above the local minimum hourly wage.

D Keyword Analysis

During filtering and modality gating, TCVP also extracts short cue phrases that are used to form the final query. These phrases are not additional annotations from humans. They are intermediate keywords produced by the LLM from the timestamped comment and the corresponding visual and audio captions. We visualize these keywords to inspect what kinds of cues remain after filtering and how they differ across modality labels.

Figure 7 shows keyword clouds for visual-related, audio-related, and mixed pairs. Audio-related pairs contain many speech-centered cues, such as topics, quoted lines, discussions. Visual-related pairs contain cues tied to visible content, such as objects, and actions. Mixed pairs include both types of cues, for example laughter or music together with visible stage, crowd, or screen cues.

Figure 8: **Modality structure across video domains.** We show the modality distribution for all 20 video domains in TCVP, sorted by visual-related share. The inner ring shows audio-related, mixed, and visual-related proportions. The outer ring further decomposes each modality label by its focus type, with mixed labels represented by paired audio and visual focus types.

This pattern is consistent with the modality labels used by TCVP: audio-related queries are mostly formed from spoken or audible content, visual-related queries from visible content, and mixed queries from combinations of the two.

E Dataset Categories

Table 6 lists representative source channels for the 20 video categories used in TCVP. Figure 8 shows the modality distribution for all 20 categories, sorted from higher visual-related share to lower visual-related share.

The category distribution forms a broad continuum rather than a single dominant pattern. At the visual-heavy end, Design, Entertainment, Sports, and Travel contain many comments grounded in visible actions, scenes, and objects. At the audio-heavy end, Podcast, Finance, Talk show, and Debate are dominated by spoken topics and quoted lines. Between these extremes, Technology, Gaming, and Fitness are more balanced, with visual, audio, and mixed evidence appearing in closer proportions. This pattern shows that the evidence needed for retrieval depends strongly on video category. As a result, a pooled retrieval score can reflect category mix as much as model ability.

The fine-grained focus-type distribution shows clear structure within each modality. Among visual-related samples, the most common focus types are action (30.2%), scene (22.8%), and object (20.6%), followed by facial expression (13.0%). Among audio-related samples, spoken topic (65.1%) and spoken quote (28.2%) account for most cases, while music (3.5%) and non-speech sound (3.1%) appear much less often. Among mixed samples, the most frequent combinations are action with spoken topic (16.4%), scene with spoken topic (16.2%), and object with spoken topic (11.2%). These re-

Category	Channels
Knowledge	BigThink, Kurzgesagt, Veritasium, Vsauce
Documentary	NatGeo, Vox, WendoverProductions
Podcast	HubermanLab, LexFridman, joerogan, TimFerriss
Debate	OxfordUnion
News	PhilipDeFranco, TheYoungTurks, DWNews
Talk show	TED, JordanBPeterson, TheDailyShow
Variety show	fallontonight, MrBeast, GoodMythicalMorning
Travel	KaraandNate, AbroadinJapan, DrewBinsky
Culture	Jolly, AsianBoss, GeographyNow
Cooking	bingingwithbabish, bonappetit, GordonRamsay
Making	MarkRober, SmarterEveryDay, Corridor
Technology	mkbhd, LinusTechTips, Mrwhosetheboss
Art/design	ProkoTV, theartassignment, TheFutur
Fashion	bestdressed, HauteLeMode, Vogue
Sports	NBA, fifa, Olympics
Gaming	PewDiePie, markiplier, pokimane, LoLEsports
Comedy	SaturdayNightLive, KeyandPeele, Smosh
Finance	GrahamStephan, AndreiJikh, MeetKevin
Fitness	AthleanX, JeffNippard, ChrisHeria
Entertainment	TopGear, TheEllenShow, WatchMojo

Table 6: **Representative source channels across video domains.** We list example YouTube channels for the 20 video domains used in dataset collection and analysis.

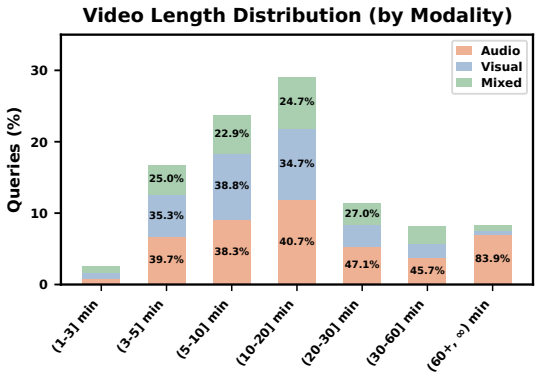

Figure 9: **Video duration distribution by modality label.** We show the distribution of retained queries across video-duration bins, excluding unrelated samples. Bar height indicates the percentage of retained queries in each duration bin, while stacked colors show the audio-related, visual-related, and mixed proportions within that bin.

sults suggest that mixed cases often pair spoken content with a visible event or entity, rather than combining two low-level perceptual cues.

The same structure appears across video domains. Design is visual-oriented, with action, object, and scene cues accounting for much of its retained set. Sports is also visual-oriented, but it includes more mixed cases involving action with spoken topic and action with music. Technology is closer to balanced, with object-centered visual evidence and spoken-topic audio evidence appearing at similar rates. Fitness shows the most balanced modality profile, and many of its mixed samples combine visible exercise actions with spoken instruction. Podcast and Finance are audio-oriented, with most retained comments grounded in spoken topics or quoted lines. Together, these patterns support modality-specific evaluation and domain-aware analysis, rather than relying only on a single pooled retrieval score.

Category descriptions. We define 20 video domains to support the domain-level analysis of viewer focus and modality-label distributions: **Design** (drawing, illustration, and visual creation), **Comedy** (sketches, stand-up, and humor), **Cooking** (recipes and kitchen demonstrations), **Culture** (cross-cultural, geographic, and social interpretation), **Debate** (formal debates and adversarial arguments), **Documentary** (factual storytelling), **Entertainment** (celebrity and show segments), **Fashion** (clothing, styling, and appearance), **Finance** (investing and markets), **Fitness** (workouts and training), **Gaming** (gameplay, esports, and streamer reactions), **Knowledge** (educational concepts and science), **Making** (building, engineering, and repair), **News** (current-events reporting), **Podcast** (long-form conversations), **Sports** (matches, highlights, and athletic action), **Talk show** (talks, interviews, and monologues), **Technology** (device reviews and product demos), **Travel** (place-based exploration), and **Variety show** (stunts, experiments).

E.1 Video duration statistics

The collected set contains 1,014 video IDs. For the duration analysis, we use a category-balanced subset so that the statistics are not dominated by a small number of frequent content types. The retained videos have a mean length of 25m 03s.

Figure 9 shows the query-level duration distribution by modality after excluding unrelated comments. Most audio, visual, and mixed queries come from videos between 5 and 20 minutes. The 5–10 and 10–20 minute bins contain the largest share of retained queries and have relatively balanced audio, visual, and mixed composition. As video length increases, the modality distribution becomes more audio-heavy: in the 60+ minute bin, 83.9% of related queries are audio-related. This trend matches the category-level statistics, where longer videos are often podcast, debate, or talk-show content. Overall, the duration analysis shows that TCVP covers both short visually grounded moments and long-form audio-centered retrieval cases.

Window	Coverage (%)
[−3s, +3s]	85
[−9s, +3s]	92
[−3s, +9s]	92
[−9s, +9s]	**97**

Table 7: **Moment coverage under different temporal windows.** We evaluate how often temporal windows of different sizes around the timestamp cover the target moment, based on 100 randomly sampled timestamped comments.

Setting	Visual	Audio	Mixed	Unrelated
TCVP (comment-grounded)	34.6%	35.0%	18.8%	11.6%
Segment-only baseline	3.9%	0.0%	93.9%	2.2%

Table 8: **Effect of timestamped comments on modality decisions.** We compare modality-label distributions with and without timestamped comments. Removing the comment causes many cases to collapse into *mixed*, while TCVP uses the comment focus to separate visual-related, audio-related, mixed, and unrelated cases.

F Pipeline Design Details

F.1 Captioning Window

Timestamped comments do not always point to the exact frame where the referred content appears. A timestamp may be placed slightly before the content becomes clear, after a spoken line, or near the moment that prompted the comment. TCVP therefore captions a short temporal window around each timestamp when generating the intermediate visual and audio captions. We use a symmetric $[-9\text{s}, +9\text{s}]$ window for modality-specific captioning. This window is long enough to include short actions, spoken phrases, and immediate reactions around the timestamp, while remaining short enough to avoid adding too much surrounding content that the comment may not refer to. To check this choice, we sample 100 timestamps and measure whether each candidate window contains the content referred to by the timestamped comment. Table 7 reports the resulting coverage across different window sizes. Extending the window to $[-9\text{s}, +9\text{s}]$ increases coverage to 97%. Larger windows add little coverage, but they increase captioning cost and can introduce additional visual or audio content unrelated to the comment. We therefore use $[-9\text{s}, +9\text{s}]$ as the captioning window in TCVP.

F.2 Is modality gating consistent?

A video segment can contain visible actions, objects, speech, and other sounds at the same time.

Modality-Specific Captioning Prompt

```
Analyze this {segment_duration}-second video
segment and write exactly two parts as two
lines.
```

`VISUAL:` Describe only what is visible, including people, objects, actions, scene context, colors, and motion.
`AUDIO:` Describe only what is audible, including speech content, speaker tone, music, sound effects, ambient sounds, and silence.

```
Keep both lines concise but informative and do
not add any other text.
```

Figure 10: **Prompt for modality-specific captioning.** Given a timestamped comment and three consecutive segment captions around the timestamp.

Although local visual and audio captions describe these cues, they do not indicate which cue the viewer referred to. To test the role of timestamped comments, we compare TCVP with a segment-only baseline that removes the comment text and assigns modality labels from only the local visual and audio captions using the same `gpt-5-mini` LLM with a matched prompt. Table 8 shows that removing the comment text shifts most labels to *mixed*, because local captions often contain both visible and audible cues even when the viewer refers to only one of them. This shows that timestamped comments help identify which part of the local evidence should guide query generation.

G Generation Prompts

We use three prompt types in the current pipeline: modality-specific captioning, comment filtering and modality gating, and query generation.

Modality-specific captioning. For each retained comment, we define an 18-second temporal window centered on the comment timestamp. To keep the captioning step local while still covering the full moment, we divide this window into three consecutive 6-second sub-segments. This gives a symmetric context around the timestamp, with one sub-segment before the center, one around the center, and one after it. We then apply modality-specific captioning to each 6-second sub-segment separately. For every sub-segment, the model produces two lines: a **VISUAL** line that describes only what is visible, and an **AUDIO** line that describes only what is audible. The visual line focuses on people, objects, actions, scene context, colors, and motion, while the audio line focuses on speech content, speaker tone, music, sound effects, ambient sounds, and silence. We concatenate the resulting captions

Comment Filtering & Modality Gating Prompt

You are helping build a research pipeline for comment-grounded video moment retrieval.

Input

- a YouTube comment
- three consecutive raw segment captions around the comment timestamp

Task

- Identify the user's main focus.
- Keep only the part supported by the segment captions.
- Assign one grounded modality gate: `visual`, `audio`, `mixed`, or `unrelated`.
- Do not generate retrieval queries in this step.

Rules

- Use only evidence grounded in the comment and the provided segment captions.
- Do not invent dialogue, sounds, objects, text, or actions not supported by the captions.
- The comment determines the user focus.
- If the comment is noisy, generic, meme-like, or poorly grounded, state that explicitly.
- Quoted text in the comment is not enough to justify audio evidence unless the captions support spoken words, singing, narration, or a distinctive sound event.
- Generic statements such as "someone is speaking" are not strong audio evidence.
- Treat generic visual framing as weak visual evidence.
- Use `mixed` only when both visual and audio evidence are individually distinctive and both are needed for retrieval.
- If the comment is not well grounded, set `primary_modality_hint` to `unrelated`.

Return valid JSON only.

```
{
"comment_focus_summary": "what the commenter seems to care about most",
"comment_focus_keywords": ["2-5 short focus keywords"],
"integrated_visual": "brief dense visual summary from the captions",
"integrated_audio": "brief dense audio summary from the captions",
"grounded_visual_evidence": "short grounded visual evidence summary",
"grounded_audio_evidence": "short grounded audio evidence summary",
"primary_modality_hint": "visual | audio | mixed | unrelated",
"audio_subtype": "speech | non_speech_audio | music | none",
"visual_focus": "action | facial_expression | object | scene | text_on_screen | interaction | none",
"audio_focus": "spoken_quote | spoken_topic | music | non_speech_sound | crowd_reaction | none",
"focus_target": "short grounded phrase naming the main cue",
"modality_gate_reason": "brief explanation of why the chosen modality gate fits the user's grounded focus",
"notes": "short note about weak evidence, unsupported details, or ambiguity"
}
```

Figure 11: **Prompt for comment filtering and modality gating.** Given a timestamped comment and three consecutive segment captions around the timestamp, the model identifies the grounded comment focus and determines whether the focus is supported by visual evidence, audio evidence, both, or neither. The output is used to retain grounded samples and assign a coarse modality label.

across the three sub-segments to form the visual and audio evidence for the full 18-second window.

Modality gating. Given a timestamped user comment and three consecutive segment captions around the timestamp, the model first identifies the main focus of the comment. It then removes unsupported or unrelated details and keeps only the evidence grounded in the captions. Using this grounded evidence, the model assigns one coarse modality label among `visual`, `audio`, `mixed`, and `unrelated`. The output JSON also records finer-grained evidence attributes. For the visual side, the model assigns one visual focus type among `action`, `facial_expression`, `object`, `scene`, `text_on_screen`, `interaction`, and `none`. For the audio side, it assigns one audio subtype among `speech`, `non_speech_audio`, `music`, and `none`, and one audio focus type among `spoken_quote`, `spoken_topic`, `music`, `non_speech_sound`, `crowd_reaction`, and `none`.

The JSON further stores the grounded visual evidence, grounded audio evidence, comment focus summary, focus keywords, focus target, and a short reason for the modality decision. The prompt is organized around three steps: identifying the user's focus, grounding that focus in the segment captions, and assigning a coarse modality gate with fine-grained evidence labels. The full prompt is shown in Figure 11.

Query Generation from Gating Output Prompt

You are generating query fields for video moment retrieval from a retained timestamped comment and its modality gating output.

Input

- the original YouTube comment
- the modality label: visual, audio, mixed, or unrelated
- the comment focus
- grounded visual caption
- grounded audio caption
- retrieval keywords

Rules

- Preserve the grounded comment focus as the main cue.
- Use only the grounded evidence provided by the gating output.
- Remove timestamps, usernames, emojis, filler, and unsupported details.
- Do not invent dialogue, sounds, objects, text, actions, or context.
- Write specific, timestamp-free queries that can be used to retrieve the target moment.

Field rules

- If the label is visual, fill `visual_query` and `final_query` using grounded visual evidence.
- If the label is audio, fill `audio_query` and `final_query` using grounded audio evidence.
- If the label is mixed, use both grounded visual and grounded audio evidence in the corresponding fields and in `final_query` when both are needed.

Return valid JSON only.

```
{ "visual_query": "query using grounded visual evidence, or empty string",
"audio_query": "query using grounded audio evidence, or empty string",
"final_query": "query for retrieving the target moment, or empty string if unrelated",
"used_visual_support": true or false,
"used_audio_support": true or false,
"query_keywords": ["2-6 grounded cue phrases used in the final query"],
"query_reason": "brief reason for how the query follows the modality label and grounded evidence" }
```

Figure 12: **Prompt for comment-guided query generation.** Given the grounded comment focus, supporting evidence, and modality label from the gating stage, the model generates retrieval queries that preserve the viewer-indicated focus. The output is returned in a structured JSON format.

Query generation. Given the modality gating output, the model generates retrieval query fields only for retained samples whose modality label is `visual`, `audio`, or `mixed`. If the label is `visual`, the query is based on grounded visual caption. If the label is `audio`, the query is based on grounded audio caption. If the label is `mixed`, the query combines grounded visual and grounded audio caption when both are needed to identify the target moment. Samples labeled as `unrelated` are filtered out before this step and do not produce retrieval queries. The output JSON records the generated query fields together with the grounded caption and modality decision, which supports step-by-step inspection and error analysis.

The prompt is organized around the grounded caption from modality gating, a small set of grounding constraints, and modality-dependent rules for filling the query fields. The full query-generation prompt is shown in Figure 12.

H Human Evaluation

We conduct human evaluation with three recruited annotators recruited through institutional channels. The evaluation covers three tasks: moment preference, modality-label validation, and query preference. Before annotation, participants receive written instructions and are informed that their judgments will be used for research evaluation, reported only in aggregate, and not linked to personal identifiers. Participation is voluntary, and annotators may stop the task at any time. Annotators complete each task independently. Each item is evaluated by three annotators, and judgments are aggregated by majority vote unless otherwise specified. Annotators are compensated based on the estimated task duration, with an estimated hourly rate of at least USD 7.25.

H.1 Modality-Label Validation

This task evaluates whether the modality label assigned by TCVP is appropriate for the timestamped comment and the corresponding video content. We

LLM Judge Instructions: Modality-Label Validation

You are an evaluator for a video moment retrieval dataset. A previous model assigned one evidence-modality label to a timestamped viewer comment. Your task is to decide whether the proposed label is supported by the comment and local evidence. This is a yes/no label-support check, not a new open-ended labeling task.

Inputs: `comment`, `visual_caption`, `audio_caption`, and `proposed_label`.

Labels: `visual` is grounded mainly in visible evidence such as actions, objects, scenes, faces, or on-screen text. `audio` is grounded mainly in audible evidence such as speech, quoted lines, voice, music, laughter, or sound effects. `mixed` requires both visible and audible evidence; do not select it only because both modalities are present. `unrelated` is used when the comment is not grounded in the local moment or cannot be supported by the captions.

Rules: Answer `yes` if the proposed label is supported by the comment and local evidence, and `no` if it is clearly inconsistent. Use only the comment and captions. Do not reject a label just because another label is also possible. If the answer is `no`, provide the best alternative label.

Output schema: Return valid JSON only.

```
{
  "supported": "yes or no",
  "better_label": "visual, audio, mixed, unrelated, or none",
  "confidence": 0.0,
  "rationale": "one short sentence"
}
```

Figure 13: **LLM judge instruction for modality-label validation.** LLM judges receive the timestamped comment, local visual/audio captions, and the TCVP label, then decide whether the label is supported.

Annotator Instructions: Modality-Label Validation

1. You will see a short YouTube clip, its timestamped comment, and a predicted modality label.
2. Judge whether the label matches the evidence needed to understand or retrieve the moment referred to by the comment.
3. Labels are `visual-related`, `audio-related`, `mixed`, and `unrelated`.
4. Choose `Agree` or `Disagree`.
5. Focus on what the comment refers to, not on every visual or audio signal that appears in the clip.

Figure 14: **Survey instruction for modality-label validation.** Annotators view a short video clip around the timestamped comment and select whether the comment is grounded in visual evidence, audio evidence, both, or neither.

construct a validation subset of 100 instances by sampling 25 instances from each of the four modality labels: visual-related, audio-related, mixed, and unrelated. Each instance contains a timestamped comment, the modality label assigned by TCVP, and information about the video content around the timestamp. For human validation, annotators are shown the timestamped comment, the TCVP label, and the video clip around the timestamp. The clip covers a $[-9\text{s}, +9\text{s}]$ window centered on the timestamp. Annotators first judge whether the comment refers to visual content, audio content, both, or neither in the clip, and then indicate whether they agree with the TCVP label. For LLM validation, each judge receives the timestamped comment, the TCVP label, and the visual/audio captions used by the pipeline. The LLM judge performs a binary label-support check rather than assigning a new label. We use GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro as independent LLM judges. For both human and LLM validation, the final judgment is binary accept/reject. Human judgments are aggregated by majority vote across the three annotators. This task validates both retained labels and unrelated labels, thereby checking the modality-labeling and filtering decisions made by TCVP.

For the LLM judge validation, each item contains the timestamped comment, the visual/audio captions, and the TCVP label. The prompt asks the judge to perform a binary label-support check rather than assign a new label.

Annotator Instructions: Moment Selection

1. You will see two candidate moments from the same YouTube video.
2. Imagine that you watched the video before and later wanted to find one of these moments again using a text query.
3. Choose the moment that you would be more likely to search for later.
4. Prefer a moment that is specific, memorable, and easy to describe in text, rather than one that is generic or routine.
5. Do not choose a moment only because it contains more content.

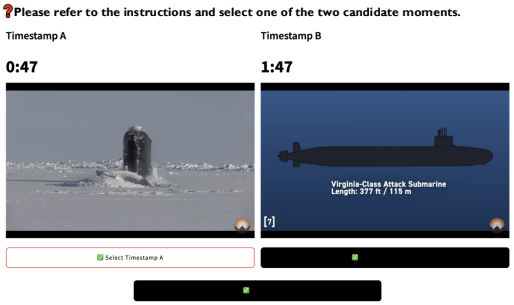

Figure 15: **Survey interface for moment preference evaluation.** The top panel shows the annotator instruction, and the bottom panel shows the survey UI. Annotators compare two candidate moments from the same video and choose which moment they would be more likely to search for later using a text query.

H.2 Moment Preference

This task evaluates whether moments indicated by timestamped comments are plausible moments that users may want to find again using text queries. Annotators are shown two clips from the same YouTube video. One clip is centered on a timestamped comment, and the other is sampled from a different timestamp in the same video. The comparison clip is chosen so that it does not overlap with the timestamp-indicated clip. Using clips from

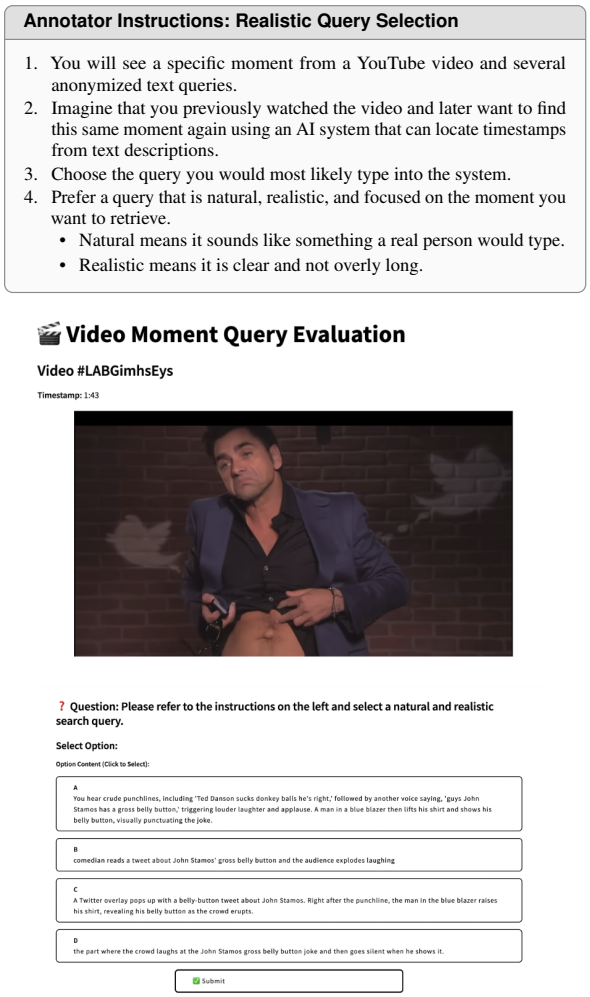

Figure 16: **Survey interface for query preference evaluation.** The figure shows the annotator instruction and the corresponding survey UI. Annotators view a target moment from a video and select the query.

the same video controls for video-level differences such as topic, style, speaker, and editing. Annotators are asked to imagine that they watched the video before and later wanted to find one of the moments again using a text query. They choose the moment they would be more likely to search for later.

H.3 Query Preference

This task compares queries generated for the same target moment. Annotators are shown the target moment and a list of candidate queries, then choose the query they would most likely use to find the moment again. The main comparison includes queries from LongVALE (Geng et al., 2025), Watch&Listen (Li et al., 2026), and TCVP. For cases where annotators select the TCVP query, we further compare TCVP queries generated with and without the timestamped comment to examine the effect of the comment signal.

Effect of Timestamped Comments in query generation. To isolate the role of the comment text after modality gating, we compare Ours w/o comments and Ours w/ comments. Both variants use the same target moment and modality label, but only Ours w/ comments uses the original timestamped comment during query generation. Among all query preference decisions, Ours w/ comments is selected in 66% of cases, while Ours w/o comments is selected in 26% of cases. This suggests that, even after the modality label is fixed, the original comment text provides additional information.

H.4 Aggregation and Replacement

Preference judgments are aggregated by majority vote. For modality-label validation, an item is accepted when at least two annotators mark the TCVP label as valid. If an item selected for downstream preference evaluation does not satisfy this validity criterion, it is replaced with another randomly sampled item under the same condition.

I Evaluation Setting & Prompts

We run all local evaluation jobs in an inference-only setting on a single machine with one NVIDIA A100 40GB GPU. The local evaluation budget is approximately 4.5 A100 GPU-hours, including segment captioning for open-source MLLMs and embedding-based retrieval. We also use API-based models for direct timestamp prediction and judge-based evaluation, and the total API cost is under 30 USD. For open-source models, parameter counts are reported in Table 5 when available. For proprietary API-based models, parameter counts are not publicly disclosed.

For segment-caption retrieval, each video is divided into non-overlapping 10-second segments. We use Qwen2.5-VL, InternVL3, and Qwen2.5-Omni to generate a short text description for each segment. For Qwen2.5-VL and InternVL3, we provide sampled visual frames from the segment. For Qwen2.5-Omni, we provide both sampled visual frames and raw audio, and instruct the model to use both visual and audio cues. In the subtitle setting, we use subtitle tracks provided by YouTube. We align subtitle text to each 10-second segment by timestamp overlap and append the aligned subtitle text to the generated segment description. This setting is used only for rows marked with S in Table 5. The query and segment descriptions are then

embedded with Qwen-3 Embedding, and segments are ranked by cosine similarity.

We use two prompt templates during evaluation. The first prompt is used for segment captioning with Qwen2.5-VL, InternVL3, and Qwen2.5-Omni. We provide a time-bounded segment and sampled frames, and ask the model to write a short caption grounded in the segment. For Qwen2.5-Omni, the prompt also includes audio input and instructs the model to incorporate audible cues when they are relevant. The second prompt is used for direct timestamp prediction. We provide the query and full-video inputs matching the modality setting in Table 5, such as sampled frames for visual-only input and sampled frames with raw audio for audio-enabled input. The model is instructed to output a single timestamp that best matches the query under a strict output format.

Segment Captioning Prompt

```
You are given a video segment #{segment_idx} from
      {start:.2f}s to {end:.2f}s.
{len(frame_timestamps)} frames are uniformly sampled at
      these timestamps (sec): [{ts_str}].

Provide a concise caption summarizing what happens in
      this segment.
Use both visual frames and the accompanying audio.
Be specific to the segment content and keep it short in
      one or two sentences.
```

Timestamp Prediction Prompt

```
Video 1 lasts for {duration_text}, and {num_frames}
      frames are uniformly sampled from it.
{timestamp_block}

Identify the single most relevant moment as one
      timestamp in seconds that matches the query.

Strict output format:
Return only one timestamp in seconds as a float.
Preferred: a bare JSON number such as 12.345.
Also acceptable: {"timestamp":12.345} or [12.345] with
      only one value.
Do not return ranges or natural language.
Keep the timestamp within [0, video_duration].

Textual query: {query}
```

All experiments in this paper evaluate existing models in an inference-only setting. We do not train or fine-tune models on TCVP. Future supervised use should split the dataset at the video level to avoid placing moment–query pairs from the same video in both training and evaluation sets.