Localizing Step-by-Step: Multimodal Long Video Temporal Grounding with LLM

Houlun Chen[1], Xin Wang[1,2,*], Hong Chen[1], Wei Feng[1], Zihan Song[1], Jia Jia[1,2,*], Wenwu Zhu[1,2,*]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University
{chenhl23,h-chen20,fw22,songzh23}@mails.tsinghua.edu.cn,{xin_wang,jjia,wwzhu}@tsinghua.edu.cn

***Abstract*—Video Temporal Grounding (VTG) localizes moments in untrimmed videos using natural language queries. Most VTG datasets focus on short videos, and existing approaches excel in short-term cross-modal matching but struggle with long VTG, where long-range temporal reasoning is required for complex events. Existing approaches typically output timestamp predictions without intermediate steps, limiting effective reasoning, whereas humans solve this step by step. To address this, we propose a long VTG framework, StepVTG, with multimodal visual and speech inputs, leveraging Large Language Models (LLMs) for step-by-step reasoning. Specifically, we transform task descriptions, speech, and visual inputs into text prompts. To enhance temporal reasoning, we introduce the Boundary-Perceptive Prompting strategy, which includes: i) a multiscale denoising Chain-of-Thought (CoT) combining global and local semantics with noise filtering, ii) validity principles to ensure LLMs generate reasonable, parsable predictions, and iii) one-shot In-Context Learning (ICL) to improve reasoning via imitation. For evaluation, we establish MM-LVTG, a new long VTG benchmark with multimodal inputs, and demonstrate through extensive experiments that StepVTG achieves state-of-the-art performance. It offers explainable reasoning steps for predictions and reveals potential in facilitating video understanding with off-the-shelf LLMs.**

***Index Terms*—Video Temporal Grounding, Long Videos, LLM**

I. Introduction

Video Temporal Grounding (VTG) [1], [2] aims to localize moments in untrimmed videos corresponding to a given query, requiring video-query context understanding and precise temporal boundary identification, as shown in Figure 1(a).

Although existing approaches [**?**], [2]–[10] have advanced VTG, they primarily focus on short videos (*e.g.,* <5 minutes [1], [2], [11]) with low-level queries and struggle with reasoning in long videos (>10 minutes [12]–[14]), common in movies, news, and courses. Long VTG requires reasoning across multiple high-level events to infer temporal boundaries over extended durations since an event is usually indirectly summarized from rich detailed objects and activities. Existing methods [4], [8], [15] often produce direct timestamps without intermediate reasoning steps, while humans solve complex tasks via decomposition. Refer to supplementary materials for detailed related works.

*Corresponding authors: Xin Wang, Jia Jia and Wenwu Zhu.

To address this, we reformulate long VTG as a long-text task, leveraging LLMs with chain-of-thought (CoT) reasoning over 10K tokens to grasp temporal boundaries, thanks to the demonstrated efficacy of LLMs in multimodal applications [16]. Specifically, we propose *StepVTG*, a versatile framework integrating off-the-shelf models via LLM prompting with speech and visual information (Figure 1(b)) for stepwise long VTG. Multimodal information is utilized as long videos often include rich speech content. First, we transcribe speeches and caption the speech- and scene-aligned frames to generate text input retaining temporal information for VTG. Subsequent experiments demonstrate such textual representation retains crucial localization information. To enhance temporal reasoning, we introduce a Boundary-Perceptive Prompting strategy with: i) multiscale denoising Chain-of-Thought (CoT) that combines global and local semantics with noise filtering, ii) validity principles for reasonable and structured predictions, and iii) one-shot in-context learning (ICL) to improve VTG task comprehension and temporal reasoning.

We validate our approach first with preliminary experiments confirming the feasibility of moment retrieval from textualized long videos. Then we establish MM-LVTG, a long VTG benchmark with multimodal inputs via collecting videos from [17], showing StepVTG achieves state-of-the-art performance over baselines. Ablation studies highlight that StepVTG benefits from both textual speech and visual modalities when handling noisy long contexts around 10K tokens, and qualitative analysis illustrates the reasoning process with explainable timestamp predictions.

We first solve long VTG with textual speech and visual modalities inputs with LLM and design a versatile training-free StepVTG framework.

To conclude, our contributions are as follows:

1) Introduce StepVTG, the first framework for long VTG using LLMs with textual speech and visual modalities in a training-free setup.
2) Propose a Boundary-Perceptive Prompting strategy enabling stepwise temporal reasoning in noisy long contexts with explainable predictions.
3) Establish MM-LVTG, a multimodal long VTG benchmark with speech and visual inputs, and demonstrate StepVTG's superiority through extensive experiments.

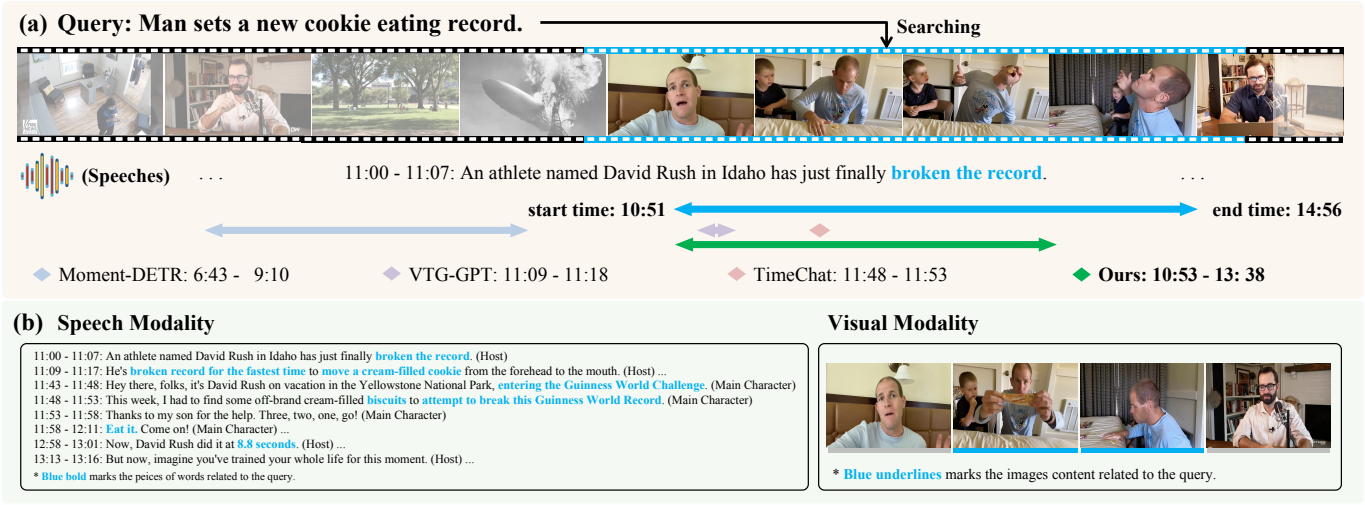

Fig. 1: This shows a 14-minute 57-second news-style video. We solve long VTG with visual and speech inputs and beat other baselines significantly.

II. StepVTG

A. Problem Formulation

Given a video V and a query Q, the method is required to predict the start-and-end timestamps of video moments $(\hat{t}_s, \hat{t}_e)$ (second) that matches query Q. For our solution, the video V with N frames is modeled as textualized representations, *i.e.* speech transcriptions $\{(t^{(i)}_{s(s)}, t^{(i)}_{s(e)}, s^{(i)})\}_{i=1}^{N_s}$ and visual captions $\{(t^{(i)}_v, c^{(i)})\}_{i=1}^{N_c}$, where N_s, N_c are the number of transcriptions and captions, respectively; the i-th piece of transcription $s^{(i)}$ lies from time $t^{(i)}_{s(s)}$ to $t^{(i)}_{s(e)}$ and the i-th piece of caption is generated on the frame sampled at time $t^{(i)}_v$. Figure 2 shows our framework.

B. Task Textualized Representations

To adapt the VTG task and multimodal inputs for LLMs, we design the following pipeline. First, we align the LLM's behavior with VTG by explaining the task and defining input-output formats (Section II-D(1)). Next, to help LLMs interpret multimodal inputs (Section II-D(2-4)), we textualize speeches and visual modalities into transcriptions and captions with temporal markers. This process retains sufficient semantics for localization and efficiently represents long videos, as validated in our experiments (Section III).

Speeches are transcribed into non-overlapping sentences using Automatic Speech Recognition (ASR). To reduce redundancy in long video frames, we employ a sampling strategy: scene changes are pre-detected, and frames are sampled in alignment with scenes and speeches. For each transcription and scene, we select the intermediate frame, *i.e.* $(t^{(i)}_{s(s)} + t^{(i)}_{s(e)})/2$ for the i-th transcription and the same for each scene, as synchronized visual contents and speeches are complementary. Captions are generated on these sampled frames. Finally, transcriptions and captions, along with query text, are provided to the LLM as task inputs. While captions can be noisy, they still improve moment localization, as demonstrated in our experiments.

C. Boundary-Perceptive Prompting

Mainstream LLM tasks rarely address temporal boundary perception in complex long contexts, making accurate VTG challenging. Moreover, LLMs' free-form responses can result in unreasonable or incomplete predictions. To tackle these issues, we propose a Boundary-Perceptive Prompting strategy, which includes a Multiscale Denoising Chain-of-Thought (CoT) for step-by-step reasoning, validity principles to regularize predictions, and one-shot In-Context Learning (ICL) to leverage LLMs' few-shot learning ability.

1) Multiscale Denoising Chain-of-Thought: We decompose Multiscale Denoising CoT into the following steps: **Step 1: Global Understanding.** We ask LLM to summarize the entire video (Section II-D(5)) to reduce detailed redundancy while preserving global high-level semantics. **Step 2: Noise Evaluation.** We ask LLM to assess how captions contribute to moment localization and adaptively balance visual and speech information gaps (Section II-D(6)). **Step 3: Partition Understanding.** We ask LLM to partition the video based on timestamps to predict and summarize each segment conditioned on the query to capture query-relevant differences among these parts (Section II-D(7)). **Step 4: Prediction.** Finally, we ask LLM to predict timestamps $(\hat{t}_s, \hat{t}_e)$ using the reasoning from prior steps.

2) Validity Principles: We define three validity principles to ensure reasonable and parsable predictions: **Format Compliance.** LLMs use a JSON template for structured, parsable outputs (Section II-D(8)). **Answer Regularization.** Predictions are constrained to ensure logical validity, *e.g.*, $\hat{t}_s < \hat{t}_e$ and one moment per query (Section II-D(9)). **Plagiarism Prohibition.** LLMs should imitate the reasoning process and format rather than copying example predictions (Section II-D(10)).

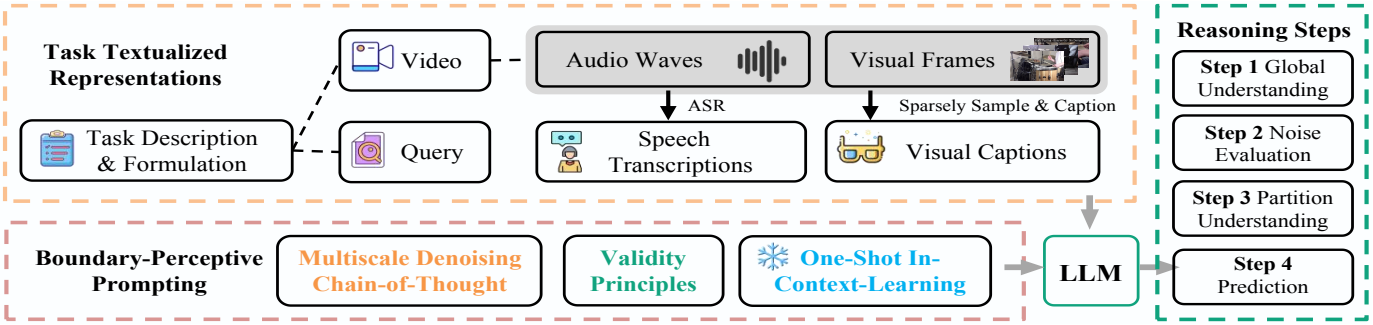

Fig. 2: Framework of our StepVTG. The task description and its speech and visual inputs are transformed into text prompts to feed LLM. To enhance the temporal reasoning capability, a Boundary-Perceptive Prompting strategy is proposed, better guiding the LLM to localize moments step by step and offering explainability.

3) One-Shot In-Context Learning: Providing a single example significantly enhances temporal reasoning and format compliance (Section II-D(11)).

D. Prompt Example

We present a prompt example in this section to better explain the details of our prompt design.

(1) Task Description & Formulation: You can analyze the correlations between a video and query, and locate the video segment that matches the query. You are given: (1) Video title (2) Query (3) Speech transcription, with temporal information in the format of: [START-TIMESTAMP] - [END-TIMESTAMP] : [TRANSCRIPTION] (4) Visual caption, with temporal information in the format of: [TIMESTAMP] : [CAPTION]. You should give the answer in [X, Y] format where X, Y are the start and end timestamps of the matching segment.

(2) Query: Habit 2: Build other people up

(3) Speech Transcriptions: 0-7: While watching clips from my last Game of Thrones video...

(4) Visual Captions: 5: A woman with long blonde hair...

(5) Global Understanding: You summarize the video.

(6) Noise Evaluation: We note that the visual caption might be quite NOISY. Now you comment if the visual captions are helpful enough for localization. You can give up information from captions if you think some of them are wrong.

(7) Partition Understanding: You analyze the video content before X, between X and Y, and after Y, respectively. After that, you give the answer [X, Y].

(8) Format Compliance: Please use JSON format of {"summary":"..." (you summarize the whole video), "comment": "..." (you evaluate effectiveness of visual captions), "query":"..." (the query input), "before X": "..." (you summarize video before X), "between X and Y": "..." (you summarize video between X and Y), "after Y": "..." (you summarize video after Y), "answer": [X, Y]}.

(9) Answer Regularization: We ensure there does exist ONE moment matching the query and X is no more than Y.

(10) Plagiarism Prohibition: You MUST NOT just copy the answer given by the example! X and Y should be replaced by the real start and end timestamps of the moment you find in videos.

(11) One-Shot In-Context-Learning: <INPUT>=> ...Query: Habit 2: Build other people up. Speech transcriptions: 0-7: While watching... Visual captions: 5: A woman... <OUTPUT>=> {"summary": "The video discusses...", "comment": "These captions describe a scene where people talk in a show, but provide limited information to understand the video.", "query": "Habit 2: Build other people up", "before 179": "Talk about...", "between 179 and 329": "Talk about...", "after 329": "Talk about...", "answer": [179, 329]}. Now you solve the following. <INPUT>=> ... <OUTPUT>=>

III. Experiments

A. Benchmark and Implementation

1) MM-LVTG Benchmark: Considering that MAD [18], a long VTG dataset, only provides extracted features without public raw videos, we create MM-LVTG using publicly available videos. We source long videos from VidChapters-7M [17], which contains 817K YouTube videos with 7M user-generated chapter annotations. These annotations summarize multiple activities, requiring methods to capture long-term relationships. We collect 618 videos (average 14 minutes each) with speech transcriptions and visual captions, sampling three query-moment pairs per video, resulting in 1,830 pairs. This forms our evaluation benchmark for multimodal long VTG, MM-LVTG (Table I). Traditional VTG benchmarks are also evaluated in supplementary materials.

2) Evaluation Metrics: We adopt the metrics "r@$\{m\}$"(%), "mIoU"(%), and "r@$\{n\}$s"(%) widely used in VTG. Here, r@$\{m\}$ measures the percentage of predictions with an IoU exceeding threshold m, while "mIoU" evaluates average localization accuracy. To address r@$\{m\}$'s limitations with lengthy ground truth moments, we use r@$\{n\}$s [17], which calculates the percentage of predictions where the predicted start-time is within n seconds of the ground truth. For LLM-based methods, if no parsable answer is generated, IoU is set to 0.0 and the start-time error to +inf.

3) Implementation Details: We use GPT-3.5-turbo-16k as the LLM model. Video scenes are detected using PySceneDetect, and sampled frames are captioned with BLIP [19]. Speech transcriptions are generated via Whisper-based tools [20]

TABLE I: Statistical Comparison between our collected MM-LVTG and other short VTG benchmarks [1], [2], [11], where V, A, and S denote visual, audio, and speech modalities, respectively. MM-LVTG focuses on evaluating the algorithm's comprehensive understanding of visual and speech content in longer videos.

Dataset	DiDeMo	Charades-STA	ActivityNet Captions	TACoS	MM-LVTG
Ave Duration (min)	0.49	0.51	1.96	4.78	13.96
Modalities	V+A	V+A	V+A	V	V+A+S

following [17]. The LLM temperature is set to 0.0 in all experiments for reproducibility. Our StepVTG serves as a flexible framework, with this being one implementation to evaluate its effectiveness. Supplementary materials include additional implementations with advanced tools (e.g., LLaVA [21], Moonshot, GPT-4o) to demonstrate its generality.

B. Baselines

To validate the contributions of visual captions and speech transcriptions in VTG, we design a rule-based BERT-based [22] baseline. Additionally, we survey related VTG technologies for long videos, categorizing them into tool-based pipelines, video multimodal LLMs, and traditional pretrained VTG models, and establish baselines accordingly. As our method is tuning-free, all evaluations are conducted under identical settings.

1) Preliminaries: We implement the following methods: (i) **Random**: Randomly select $(\hat{t}_s, \hat{t}_e)$ with $\hat{t}_s < \hat{t}_e$, reporting average metrics over 10 repetitions. (ii) **Complete**: Use the entire video duration $(0, T)$, where T is the video length. (iii) **BERT-X** [22]: Generate embeddings for transcriptions, captions, and queries using BERT. Transcriptions and captions at the same time are concatenated when both are used. Prediction is rule-based, selecting $(\hat{t}_s, \hat{t}_e)$ where the highest matching score is at $\hat{t}_s$ and decreases by ϵ at $\hat{t}_e$ ($\epsilon = 0.05$ as in [17]). Here, "X" includes {Asr, Cap, Asr+Cap}.

2) Tool-Based Pipelines: We reproduce VTG-GPT [23], a tuning-free zero-shot VTG method that employs a propose-and-match pipeline with tools like language models, caption models, and proposal generators. While the official VTG-GPT uses only video captions, we adapt it to the Asr, Cap, and Asr+Cap settings.

3) Video Multimodal LLMs: We compare our method with VideoChat [24], Video-ChatGPT [25], Video-LLaMA [26], TimeChat [8], VTG-LLM [15], and Grounded-VideoLLM [27]. For a fair comparison, all methods use video frames and speech transcriptions, with Video-LLaMA also analyzing raw audio. TimeChat, VTG-LLM, and Grounded-VideoLLM, the latest multimodal LLMs, claim explicit temporal understanding. Predictions are extracted by parsing their free-form outputs with regular expressions. We calculate the failure rate to predict a pair of start-and-end timestamps, assessing their understanding of the VTG task and output quality.

4) Traditional VTG Models: We use strong generalizable pretrained VTG models, Moment-DETR [5], R2-Tuning [28], and UniVTG [4], from their official repositories as baselines.

C. Empirical Results

Table II shows the benchmarked performances on MM-LVTG, demonstrating the feasibility of retrieving moments using textualized video representations. Our method achieves state-of-the-art results, validating its effectiveness for VTG in long videos.

1) Preliminaries: A simple BERT-based matching pipeline outperforms "random" with either speeches or captions (rows 3-5), showing that both modalities provide key query-related clues. Further, incorporating both modalities results in better results on IoU-related metrics, indicating the compatibility of both modalities.

2) Related Technologies: Existing GPT-involved methods, whether tool-based ones (rows 6-8) or most multimodal video LLMs (rows 9-14), struggle to align themselves to the VTG task due to weak perception of temporal boundaries, even underperforming on IoU-related metrics compared to random selection. VideoChat, Video-ChatGPT, and Video-LLaMA even struggle to generate a span-like answers for the VTG task. TimeChat, VTG-LLM, and Grounded-VideoLLM show better start-time localization but fail to handle moment spans effectively. New multimodal LLMs underperform tool-based VTG-GPT, underscoring the challenge of enabling video LLMs to perform complex temporal reasoning beyond basic content signal perception. Moment-DETR, R2-Tuning, and UniVTG (rows 15-17) show a strong advantage over video LLMs in IoU-related metrics but show weakness in start-time predictions. Despite pretraining, they demonstrate limited generalization in long-video scenarios, indicating room for improving pretraining from strategies or data to handle long contexts. In contrast, our tuning-free method achieves state-of-the-art results against all baselines by great margins, especially over 10× performances on r@0.9 and r@1s metrics.

D. Ablation Analysis

We perform ablation studies (Table III) to evaluate key components of our method. Rows 3-5 remove the three steps of CoT in Section II-C1 (w/o CoT) or instruct LLM to answer without one-shot examples (w/o ICL). Rows 6-7 exclude speech transcriptions or visual captions from both the one-shot example and test sample to assess reliance on multimodal information.

1) Effect of Boundary-Perceptive Prompting: Rows 3-5 highlight the effectiveness of Boundary-Perception Prompting, achieving 40- and 125-fold gains in r@0.7 and r@0.9 when both CoT and ICL are used. CoT enhances accuracy and adds explainability. LLM struggles to predict meaningful timestamps without CoT and ICL, performing worse than

TABLE II: Comparison on MM-LVTG. A, V, and S are short for audio, visual, and speech modalities, which are the input modalities from videos at inference. The number in bottom-right $(\cdot)$ denotes the proportion that the method does not generate a parsable answer. The smaller, the better. The best and second are highlighted by **bold** and underline.

Methods	Modalities	r@0.3	r@0.5	r@0.7	r@0.9	mIoU	r@1s	r@3s	r@5s	r@10s
Random	-	13.10	5.36	1.67	0.19	10.31	0.37	0.83	1.38	2.52
Complete	-	12.62	3.77	1.04	0.16	15.94	10.11	10.16	10.27	11.04
BERT-Asr	S	16.07	6.89	3.39	0.87	13.76	1.64	3.83	5.46	7.05
BERT-Cap	V	16.28	6.89	2.73	0.87	14.06	0.44	1.48	2.79	4.64
BERT-Asr+Cap	V+S	17.43	7.38	3.72	1.15	14.61	0.55	3.55	5.25	7.05
VTG-GPT-Asr	S	5.23	2.09	0.83	0.22	6.86	8.31	11.61	14.03	19.48
VTG-GPT-Cap	V	3.07	1.21	0.55	0.27	3.58	2.63	5.10	6.64	9.65
VTG-GPT-Asr+Cap	V+S	4.86	1.69	0.38	0.00	5.93	7.27	10.77	13.50	19.51
VideoChat$_{(43.84)}$	V+S	1.20	0.38	0.05	0.00	1.48	3.61	4.27	4.76	6.24
Video-ChatGPT$_{(46.72)}$	V+S	1.20	0.60	0.05	0.05	1.21	2.57	3.01	3.55	4.32
Video-LLaMA$_{(52.81)}$	A+V+S	0.60	0.24	0.12	0.06	1.16	2.60	3.50	3.98	5.25
TimeChat	V+S	0.87	0.38	0.00	0.00	2.03	7.87	9.23	10.66	15.25
VTG-LLM	V+S	2.95	1.26	0.49	0.16	2.66	5.30	5.57	5.79	6.83
Grounded-VideoLLM	V+S	3.83	1.42	0.44	0.11	3.69	9.89	10.27	10.49	11.53
Moment-DETR	V	14.59	6.34	2.30	0.33	11.68	1.15	2.51	3.28	5.08
R2-Tuning	V	10.55	3.99	1.37	0.27	7.54	0.93	1.69	1.86	3.55
UniVTG	V	18.74	8.74	3.50	0.55	14.19	1.20	3.11	4.81	8.31
StepVTG	V+S	**34.81**	**22.95**	**14.92**	**6.28**	**26.81**	**17.60**	**25.41**	**32.02**	**39.73**

TABLE III: Ablation Studies on MM-LVTG. The best and the second are highlighted by **bold** and underline, respectively.

Methods			r@0.3	r@0.5	r@0.7	r@0.9	mIoU	r@1s	r@3s	r@5s	r@10s
Random			13.10	5.36	1.67	0.19	10.31	0.37	0.83	1.38	2.52
Complete			12.62	3.77	1.04	0.16	15.94	10.11	10.16	10.27	11.04
	CoT	ICL									
Ours	✘	✘	2.40	1.26	0.38	0.05	4.37	12.84	18.85	23.50	29.89
	✓	✘	12.73	6.99	3.33	0.98	11.56	16.39	23.61	29.62	37.27
	✘	✓	21.69	13.06	7.32	2.24	17.24	16.39	23.44	27.60	34.32
	Speech	Visual									
	✓	✘	28.25	18.42	11.37	4.70	21.84	16.78	23.93	29.34	36.12
	✘	✓	20.22	10.49	5.14	1.15	15.17	5.90	10.27	12.40	15.68
	✓	✓	**34.81**	**22.95**	**14.92**	**6.28**	**26.81**	**17.60**	**25.41**	**32.02**	**39.73**

"random" and "complete" on r@$\{m\}$ and mIoU (row 3). CoT and ICL improve temporal coverage by better aligning start and end timestamps rather than single-point localization, proving both are crucial for VTG understanding. While CoT and ICL achieve comparable r@$\{n\}$s, ICL nearly doubles r@$\{m\}$ performance, showing LLM's strength in example-driven understanding (rows 4-5).

Fig. 3: A Cook long video case.

2) Effect of Multimodal Information: Rows 6-8 demonstrate the importance of combining speech and visual modalities, significantly outperforming single-modality setups. (i) Removing speech reduces r@0.9 by 5.13%, and removing captions leads to a 1.58% drop, confirming the need for both. (ii)Start time prediction is close to the optimal only with speeches (row 6), but captions enhance IoU accuracy when combined (row 8), highlighting the importance of multimodal information for VTG in long videos. (iii) Despite noise and information loss in sparse captions, LLM benefits from visual data, leveraging Boundary-Perceptive Prompting to filter irrelevant content and extract meaningful insights for VTG.

E. Qualitative Analysis

Figure 3 showcases the reasoning process of StepVTG, which enhances the explainability of its prediction. Additional cases, along with speech and visual inputs, are provided in the supplementary materials.

a) Query: BBQ Sandwiches

b) Ground Truth: (38, 160)

c) Prediction: {"***summary***": "The video demonstrates how to make barbecued chicken sandwiches. The recipe includes using chicken breasts, Worcestershire sauce, Italian dressing, brown sugar, and barbecue sauce. The chicken is cooked in a crockpot for six hours and then shredded. The video also shows how to make coleslaw as a side dish. The coleslaw is made with pre-mixed carrots and cabbage, salad dressing, mayonnaise, sugar, and vinegar. The video concludes with the final dish of barbecued chicken sandwiches and coleslaw.", "***comment***": "The visual captions provide helpful

information about the ingredients and cooking process, making it easier to follow along with the video. They are helpful for localization.", "***query***": "BBQ Sandwiches", "***before 38***": "talk about the introduction and the quick and easy meals featured in the video.", "***between 38 and 80***": "talk about the recipe and ingredients for the barbecued chicken sandwiches.", "***between 80 and 160***": "talk about cooking the chicken in the crockpot and preparing the coleslaw.", "***after 160***": "talk about the final dish and conclude the video.", "***answer***": [38, 160] }

IV. Conclusion

We propose StepVTG, a framework for Video Temporal Grounding (VTG) in long videos using multimodal speech and visual inputs in a step-by-step, tuning-free manner. Our Boundary-Perceptive Prompting improves long-range temporal reasoning and boundary detection in noisy contexts up to 10K tokens. Addressing the limitations of existing VTG benchmarks focused on short videos and neglecting speech, we establish the MM-LVTG benchmark for long VTG with visual and speech inputs. Experiments show that StepVTG outperforms baselines while providing explainable predictions. By reformulating VTG as a long-textual task, we showcase LLMs' potential for complex temporal video tasks. While speech and visual inputs offer key clues for long VTG, they cannot fully represent videos, as captions act as lossy visual compressions. Comprehensive video understanding requires integrating more modalities and efficient representations. Future work could explore extracting task-specific information, such as guiding captioning models with text queries to focus on VTG-relevant content.

Acknowledgment

This work is supported by National Natural Science Foundation of China No.62222209, Beijing National Research Center for Information Science and Technology under Grant No.BNR2023TD03006.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell, "Localizing moments in video with natural language," in *Proceedings of the IEEE ICCV*, 2017, pp. 5803–5812.

[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE ICCV*, 2017, pp. 5267–5275.

[3] Wei Feng, Xin Wang, Hong Chen, Zeyang Zhang, Houlun Chen, Zihan Song, Yuwei Zhou, Yuekui Yang, Haiyang Wu, and Wenwu Zhu, "Llm4vg: Large language models evaluation for video grounding," *arXiv preprint arXiv:2312.14206*, 2023.

[4] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou, "Univtg: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF ICCV*, 2023, pp. 2794–2804.

[5] Jie Lei, Tamara L Berg, and Mohit Bansal, "Detecting moments and highlights in videos via natural language queries," *Advances in NeurIPS*, vol. 34, pp. 11846–11858, 2021.

[6] Houlun Chen, Xin Wang, Xiaohan Lan, Hong Chen, Xuguang Duan, Jia Jia, and Wenwu Zhu, "Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding," in *Proceedings of the 31st ACM MM*, 2023, pp. 3117–3128.

[7] Houlun Chen, Xin Wang, Hong Chen, Zeyang Zhang, Wei Feng, Bin Huang, Jia Jia, and Wenwu Zhu, "Verified: A video corpus moment retrieval benchmark for fine-grained video understanding," in *The Thirty-eight Conference on NeurIPS*.

[8] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou, "Timechat: A time-sensitive multimodal large language model for long video understanding," *arXiv preprint arXiv:2312.02051*, 2023.

[9] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu, "Vtimellm: Empower llm to grasp video moments," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2024, pp. 14271–14280.

[10] Bin Huang, Xin Wang, Hong Chen, Houlun Chen, Yaofei Wu, and Wenwu Zhu, "Identity-text video corpus grounding," 2025.

[11] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles, "Dense-captioning events in videos," in *Proceedings of the IEEE ICCV*, 2017, pp. 706–715.

[12] Chao-Yuan Wu and Philipp Krahenbuhl, "Towards long-form video understanding," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2021, pp. 1884–1894.

[13] Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid, "Selective structured state-spaces for long-form video understanding," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2023, pp. 6387–6397.

[14] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu, "Long-form video-language pre-training with multimodal temporal contrastive learning," *Advances in NeurIPS*, vol. 35, pp. 38032–38045, 2022.

[15] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Xi Chen, and Bo Zhao, "Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding," *arXiv preprint arXiv:2405.13382*, 2024.

[16] Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu, "Multi-modal generative ai: Multi-modal llm, diffusion and beyond," *arXiv preprint arXiv:2409.14993*, 2024.

[17] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid, "Vidchapters-7m: Video chapters at scale," *arXiv preprint arXiv:2309.13952*, 2023.

[18] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem, "Mad: A scalable dataset for language grounding in videos from movie audio descriptions," in *Proceedings of the IEEE/CVF Conference on CVPR*, 2022, pp. 5026–5035.

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*. PMLR, 2022, pp. 12888–12900.

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*. PMLR, 2023, pp. 28492–28518.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," *Advances in NeurIPS*, vol. 36, 2024.

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, vol. 1, p. 2.

[23] Yifang Xu, Yunzhuo Sun, Zien Xie, Benxiang Zhai, and Sidan Du, "Vtg-gpt: Tuning-free zero-shot video temporal grounding with gpt," *Applied Sciences*, vol. 14, no. 5, pp. 1894, 2024.

[24] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[25] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv preprint arXiv:2306.05424*, 2023.

[26] Hang Zhang, Xin Li, and Lidong Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[27] Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang, "Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models," *arXiv preprint arXiv:2410.03290*, 2024.

[28] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen, "R2-tuning: Efficient image-to-video transfer learning for video temporal grounding," *arXiv preprint arXiv:2404.00801*, 2024.