Mitigating the Discrepancy Between Video and Text Temporal Sequences: A Time-Perception Enhanced Video Grounding method for LLM

Xuefen Li[1]†, Bo Wang[2, 3]†, Ge Shi[1]*, Chong Feng[2, 3], Jiahao Teng[1]
[1]Beijing University of Technology, China
[2]Beijing Institute of Technology, China
[3]Southeast Academy of Information Technology, Beijing Institute of Technology, China
shige@bjut.edu.cn

Abstract

Existing video-LLMs excel at capturing the overall description of a video but lack the ability to demonstrate an understanding of temporal dynamics and a fine-grained grasp of localized content within the video. In this paper, we propose a Time-Perception Enhanced Video Grounding via Boundary Perception and Temporal Reasoning aimed at mitigating LLMs' difficulties in understanding the discrepancies between video and text temporality. Specifically, to address the inherent biases in current datasets, we design a series of boundary-perception tasks to enable LLMs to capture accurate video temporality. To tackle LLMs' insufficient understanding of temporal information, we develop specialized tasks for boundary perception and temporal relationship reasoning to deepen LLMs' perception of video temporality. Our experimental results show significant improvements across three datasets: ActivityNet, Charades, and DiDeMo (achieving up to 11.2% improvement on R@0.3), demonstrating the effectiveness of our proposed temporal awareness-enhanced data construction method.[1]

1 Introduction

With the success of large language models (LLMs) in the field of natural language processing (Touvron et al., 2023) (Achiam et al., 2023), an increasing number of researchers are attempting to leverage the capabilities of LLMs in the domain of video understanding, leading to the emergence of various video-LLMs (Ko et al., 2023). Video Grounding, as a representative task in video temporal understanding, aims to locate specific events in an untrimmed video based on natural language descriptions.

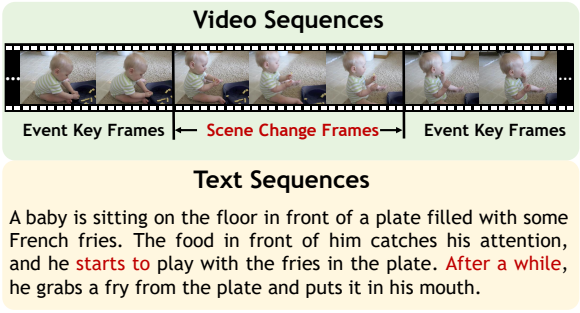


Figure 1: Illustration of video sequences and corresponding text sequences. The video frames show key events and scene changes, while the text describes the action taking place in the video. Text marked in red represents the turning point markers.

Traditional methods often require the meticulous design of complex multimodal fusion frameworks to achieve precise temporal localization (Cao et al., 2023), researchers have been inspired to explore the application of LLMs in the multimodal domain. The application of large language models (LLMs) for video grounding tasks remains in the exploratory phase. (Li et al., 2023) (Zhang et al., 2023) (Maaz et al., 2023) have shown impressive capabilities in video understanding, they reveal significant limitations in addressing fine-grained temporal questions. To address these challenges, existing approaches construct temporally contextualized conversational datasets to perform supervised fine-tuning of LLMs. (Huang et al., 2024) uses a three-stage training approach to fine-tune the video-LLM step by step. (Li et al., 2024) developed video-LLMs that leverage large-scale multimodal datasets, incorporating both video-text and audio data to enhance temporal comprehension. However, their data construction methods often simplify video content into a series of independent event descriptions, such as: *"Q: During which frames can we see T_i happening in the video? A: From t_s to t_e."* Here, T represents the event, t_s represents the start time, and t_e represents the end time. This data

*Corresponding author. †These authors contributed equally to this work.
[1]Our code will available at https://github.com/lixuefenfen/TPE-VLLM.

Proceedings of the 31st International Conference on Computational Linguistics, pages 9804–9813
January 19–24, 2025.

construction method treats each video segment in isolation during the supervised fine-tuning phase, neglecting the inter-dependencies between different events within the same video.

In terms of the natural differences between video and text modalities, as illustrated in Figure 1, the temporal relationships between video frames are more continuous and smooth, whereas the gaps between words in text can be larger and structurally more complex. Moreover, individual video frames typically contain more information than single words or characters, requiring the model to have stronger information extraction and compression capabilities. A critical examination of existing methods and modal differences reveals two primary challenges that contribute to the current gap in video grounding performance.

Inadequate Perception of Event Boundaries: Current methods lack sufficient mechanisms for accurately identifying and delineating the boundaries of events within videos. This issue arises because most models are trained to comprehend overall video content but fail to detect the precise moments when an event begins or ends. Consequently, these methods are ill-equipped to address tasks that require nuanced temporal boundary detection.
Insufficient Temporal Understanding: Beyond the inter-training limitations, there is a significant gap in how existing models handle tasks that rely on a deep understanding of temporal sequences. Most current approaches focus on coarse-grained content comprehension rather than on fine-grained temporal reasoning. This inadequacy becomes particularly evident in video grounding, where the model needs to understand the temporal relationships within the video with greater granularity.

To address the two core issues, limited event boundary perception, and insufficient temporal understanding, we proposed Time-Perception Enhanced Video Grounding via Boundary Perception and Temporal Reasoning (TPE-VLLM). Specifically, To tackle the problem of inaccurate identification and delineation of event boundaries within videos, we designed boundary perception tasks. These tasks include duration perception and position perception subtasks, which enable the model to more precisely capture the start and end points of events, thus improving its ability to handle tasks requiring nuanced temporal boundary detection. To address the lack of fine-grained temporal understanding in current models, we developed temporal reasoning tasks. These tasks include event time matching, event ordering, and time selection for given events, which enhance the model's comprehension of complex temporal relationships within videos, enabling more accurate temporal reasoning. The results show the significant improvements in video grounding performance on ActivityNet, Charades, and DiDeMo datasets. Our contributions are as follows:

- We introduced boundary perception and temporal reasoning tasks, significantly enhancing the model's event boundary detection and temporal understanding.
- Experiments demonstrated the efficiency of these tasks, achieving strong performance with just 20% of the constructed training data.
- Extensive evaluations on ActivityNet, Charades-STA, and DiDeMo showed notable improvements, validating the effectiveness of our time-perception enhanced approach.

2 Related Work

Video Grounding The video grounding task aims to locate specific segments in untrimmed videos based on text descriptions (Gao et al., 2017). Traditionally, achieving performance improvements requires carefully designed model architectures (Zhang et al., 2021). To enhance model universality, researchers are increasingly using general models for video grounding. For instance, (Zheng et al., 2023) employs BLIP (Li et al., 2022) to convert video frame content into text descriptions and declusters frames based on the differences in event descriptions. Similarly, (Luo et al., 2024), breaking down text into shorter segments, utilizes InternVideo (Wang et al., 2022) to match text with video clips. However, these sophisticated models often demand experienced engineers to design complex systems that may not be suitable for contemporary modeling environments.

LLMs for Video Grounding With the impressive performance of LLMs in NLP, researchers are increasingly exploring the application of LLMs in video grounding (Liu et al., 2024). Video LLMs directly accept videos and queries, responding based on the query and video content like VideoChat (Li et al., 2023) and Video-LLama (Zhang et al., 2023). However, these general-purpose models have shown unsatisfactory performance in video grounding. Recently, efforts have shifted toward fine-tuning LLMs for this task. (Huang et al., 2024) proposes a three-stage time sensing model

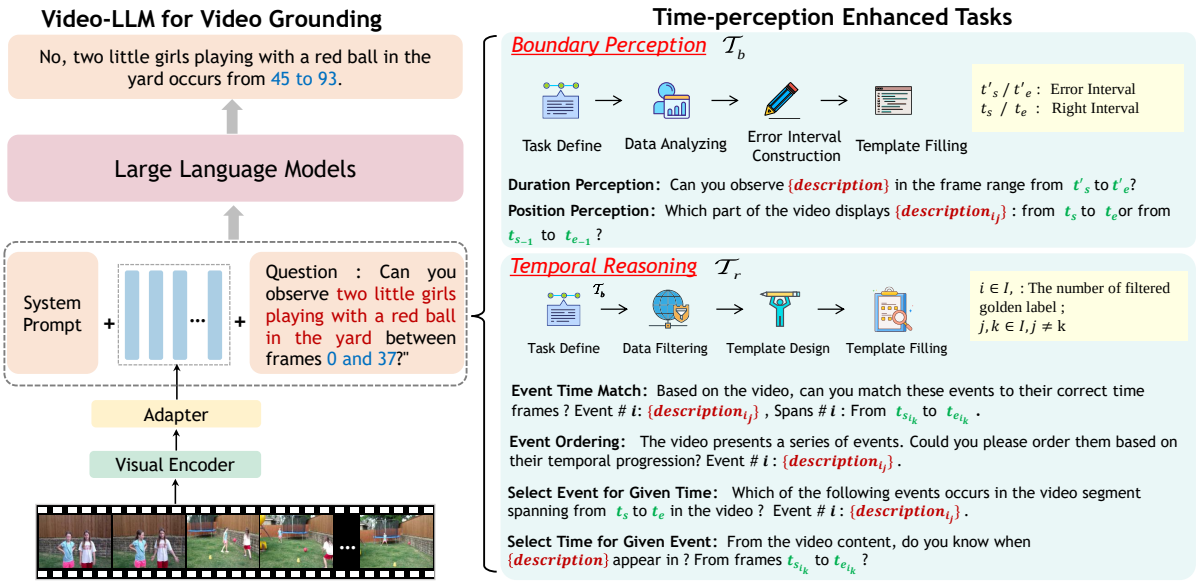


Figure 2: The illustration of TPE-VLLM: 1) Our training framework based on LLM, 2) Our method for constructing boundary-aware and temporal reasoning dialogue data. We use a high-quality dataset to construct Boundary Perception and Temporal Reasoning dialogue tasks, and merge them train Stage3 to enrich the training tasks, guide the model to correctly perceive boundaries and distinguish between different events.

to create a dialogue dataset focused on temporal perception. (Li et al., 2024) introduces the incorporation of audio modality to further enhance the model's capabilities and presents a strategy for constructing training data. However, the template form of the existing dialogue data are often singular in form and constructed based on individual events, which leads to a deficiency in the model's ability to reason about temporal relationships.

3 Approach

Our approach builds upon a multi-stage architecture, with significant innovations introduced in the third stage to enhance the model's temporal reasoning capabilities. As illustrated in Figure 2, TPE-VLLM consists of three main components: a multi-stage architecture (Section 3.1), novel time-perception enhanced tasks (Section 3.2), and a specialized data fusion and training strategy (Section 3.3). The core of our contribution lies in the time-perception enhanced tasks, which include boundary perception tasks and temporal reasoning tasks, specifically designed to improve the model's ability to perceive event boundaries accurately and understand complex temporal relationships within videos.

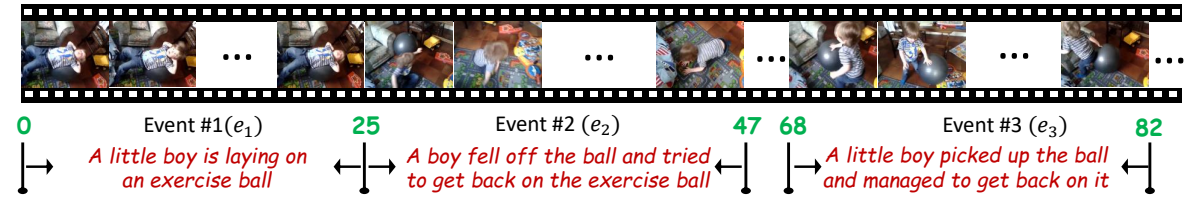
3.1 Multi-Stage Architecture

We first formally define the video grounding task. Given an untrimmed video $V = v_{t=1}^{T}$ of T frames and a natural language query Q, the goal is to locate the temporal segment $[t_s, t_e]$ that best corresponds to the query, where $1 \leq t_s < t_e \leq T$. Formally, we aim to learn a mapping function f:

$$f : (V, Q) \rightarrow [t_s, t_e] \tag{1}$$

TPE-VLLM adopts a three-stage training methodology following (Huang et al., 2024), as illustrated in Figure 2. In the first stage, we use a frozen CLIP ViT-L/14 (Radford et al., 2021) to extract visual features from uniformly sampled video frames and train a Visual Adapter to project these features into the LLM's embedding space. The second stage fine-tunes the LLM on extensive video datasets (Wang et al., 2023) using Low-Rank Adaptation (LoRA), enabling effective video content comprehension while maintaining most pre-trained parameters. The third stage represents the core innovation of our method. We leverage Video Grounding datasets to construct high-quality data, introducing two novel types of tasks: Boundary Perception Tasks $\mathcal{T}_b$ and Temporal Reasoning Tasks $\mathcal{T}_r$, which will be elaborated in Section 3.2. We perform Supervised Fine-Tuning (SFT) in this stage, using the following loss function:

$$\mathcal{L} = -\log P(y|V, Q, \mathcal{T}_b, \mathcal{T}_r) \tag{2}$$



Event Time Match (ETM):

Question : Based on the video, can you match these event s to their correct time frames? Event #1: *A little boy picked up the ball and managed to get back on it.* Event #2: *A little boy is laying on an exercise ball.* Event #3: *The boy fell off the ball and tried to get back on the exercise ball.* Span #1: From 0 to 25, Span #2: From 68 to 82. Span #3: From 25 to 47.

Answer: The correct matches are: Event #1 matches span #2, Event #2 matches span #1, Event #3 matches span #3.

Event Ordering (EO):

Question : The video presents a series of events. Could you please order them based on their temporal progression? Event #1: *A little boy picked up the ball and managed to get back on it.* Event #2: *A little boy is laying on an exercise ball.* Event #3: *The boy fell off the ball and tried to get back on the exercise ball.*

Answer: According to the video content, the sequence of events should be: Event #2, Event #1, Event #3.

Event Selection for Given Time (ES):

Question : Which of the following events occurs in the video segment spanning from 25 to 47? Event #1: *A little boy picked up the ball and managed to get back on it.* Event #2: *A little boy is laying on an exercise ball.* Event #3: *The boy fell off the ball and tried to get back on the exercise ball.*

Answer: The event that occurs in the video between 25 and 47 is *a boy fell off the ball and tried to get back on the exercise ball.* This corresponds to option event 3 in the list."

Time Selection for Given Event (TS):

Question : From the video content, do you know when *a little boy picked up the ball and managed to get back on* appear in? From frames 0 to 25, from 25 to 47 or from 68 to 82?

Answer: Based on the video content, *a little boy picked up the ball and managed to get back on* appears from 68 to 82.

Figure 3: Example of temporal video reasoning tasks, which presents a video segmented into three distinct events. Red text annotates the descriptions of events, while green markers highlight the corresponding video frames.

where y is the model's output. This loss function integrates the standard video grounding objective with our novel tasks, enabling the model to jointly optimize for accurate temporal localization and enhanced temporal understanding.

3.2 Time-Perception Enhanced Tasks

The core innovation of our approach lies in the design of specialized tasks for the third stage of training, aimed at enhancing the LLM's temporal understanding in video content. We introduce two categories of tasks: Boundary Perception Tasks and Temporal Reasoning Tasks. These tasks are designed to address two key challenges in video grounding: (1) inadequate perception of event boundaries and (2) insufficient temporal relationship understanding.

3.2.1 Boundary Perception Tasks

Boundary Perception Tasks are designed to improve the model's ability to accurately identify the start and end times of events in videos. These tasks address the common biases in existing models, which often struggle with precise temporal localization.

We construct two types of Boundary Perception Tasks: **Duration Perception Task**: This task aims to correct the model's tendency to over- or under-estimate event durations. **Position Perception Task**: This task focuses on improving the model's ability to accurately locate events within the overall video timeline. The construction of these tasks follows Algorithm 1, which generates question-answer pairs based on manipulated time intervals.

The `GenerateErrorInterval` function creates erroneous intervals by applying shifts sampled from the bias distributions B_{dur} and B_{pos}. These distributions are derived from the error patterns observed in our Stage 2 model. The GenerateQuestion function formulates a question that requires the model to distinguish between the correct and erroneous intervals. For example:

Duration Perception:

Q: *In which segment do you see {description} happening: t'_s to t'_e or t_s to t_e?*

A: *{description} occurs from t_s to t_e.*

Position Perception:

Q: *Can you observe {description} in the frame range from t'_s to t'_e?*

A: *No, The event { description} happens from t_s to t_e.*

This task design encourages the model to develop a more nuanced understanding of event durations and positions within videos.

Algorithm 1 Boundary Perception Task Generation

Require: Video dataset D, bias distributions B_{dur}, B_{pos}

Ensure: Generated task set T_{BP}

for each video $v \in D$ **do**
  Extract ground truth interval $[t_s, t_e]$ and event description e
  $[t'_s, t'_e] \leftarrow$ `GenerateErrorInterval` $([t_s, t_e], B_{dur}, B_{pos})$
  $q \leftarrow$ `GenerateQuestion`$(e, [t_s, t_e], [t'_s, t'_e])$
  $a \leftarrow$ `GenerateAnswer`$([t_s, t_e])$
  Add (q, a) to T_{BP}
end for
return T_{BP}

3.2.2 Temporal Reasoning Tasks

While Boundary Perception Tasks focus on individual events, Temporal Reasoning Tasks are designed to enhance the model's understanding of the relationships between multiple events within a video. These tasks address the challenge of comprehending the complex temporal dynamics in videos containing multiple events.

We introduce four types of Temporal Reasoning Tasks: **Event Time Matching**, which requires the model to associate events with their correct time intervals in a multi-event video; **Event Ordering**, which challenges the model to arrange events in their chronological sequence; **Event Selection for Given Time**, testing the model's ability to identify which events occur within a specified time interval; and **Time Selection for Given Event**, which requires the model to select the correct time interval for a given event description. The construction of these tasks follows Algorithm 2, which generates diverse question-answer pairs based on the temporal relationships between events.

The `MeetsFilteringCriteria` function ensures that selected videos and events meet the following criteria for task generation: video duration is between 10 and 200 seconds, event durations range from 10% to 90% of the video length, events do not overlap, and each video contains at least three events to support complex reasoning.

Algorithm 2 Temporal Reasoning Task Generation

Require: Video dataset D, task types $T = \{t_1, t_2, t_3, t_4\}$

Ensure: Generated task set T_{TR}

for each video $v \in D$ **do**
  Extract event set $E = \{(e_1, [t_{s1}, t_{e1}]), ..., (e_n, [t_{sn}, t_{en}])\}$
  if `MeetsFilteringCriteria`(v, E) **then**
    for each task type $t \in T$ **do**
      $q \leftarrow$ `GenerateQuestion`(t, E)
      $a \leftarrow$ `GenerateAnswer`(t, E)
      Add (q, a) to T_{TR}
    end for
  end if
end for
return T_{TR}

Examples of generated questions for each task type are as follows:

Event Time Matching (ETM):

Q: *Based on the video, can you match these events to their correct time frames? : Event #1: {description}, Event #2: {description}, Event #3: {description}. Span #1: $[t_{s1}, t_{e1}]$, Span #2: $[t_{s2}, t_{e2}]$, Span #3: $[t_{s3}, t_{e3}]$.*

Event Ordering (EO):

Q: *The video presents a series of events. Could you please order them based on their temporal progression? : Event #1: {description}, Event #2: {description}, Event #3: {description}.*

Event Selection for Given Time (ES):

Q: *Which of the following events occurs in the video segment spanning from t_s to t_e in the video? Event #1: {description}, Event #2: {description}, Event #3: {description}.*

Time Selection for Given Event (TS):

Q: *From the video content, do you know when "{description}" appear in? A. $[t_{s1}, t_{e1}]$, B. $[t_{s2}, t_{e2}]$, C. $[t_{s3}, t_{e3}]$.*

These tasks collectively challenge the model to develop a comprehensive understanding of temporal relationships within videos, going beyond simple start and end time identification. We show template examples of these tasks and populated examples using description and timestamps, as shown in Figure 3.

Model	ActivityNet				Charades-STA				DiDeMo			
	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
7B Series												
VideoChat-7B	8.8	3.7	1.5	7.2	9.0	3.3	1.3	6.5	-	-	-	-
VideoLLaMA-7B	6.9	2.2	0.8	6.5	10.4	3.8	0.9	7.1	-	-	-	-
VideoChatGPT-7B	26.4	13.6	6.1	18.9	20.0	7.7	1.7	13.7	-	-	-	-
GroundingGPT-7B	-	-	-	-	-	29.6	11.9	-	-	-	-	-
VTimeLLM-7B	44.0	27.8	14.3	30.4	51.0	27.5	11.4	31.2	36.3	28.8	20.9	27.9
TimeChat-7B	-	-	-	-	-	32.2	13.4	-	-	-	-	-
Ours-7B	**50.4**	**35.4**	**19.2**	**36.1**	**55.5**	**33.1**	**14.7**	**34.7**	**36.9**	**29.8**	**22.5**	**29.0**
13B Series												
VTimeLLM-13B	44.8	29.5	14.2	31.4	55.3	34.3	14.7	34.6	43.6	33.0	23.4	32.2
TPE-VLLM-13B	**55.0**	**37.1**	**20.0**	**38.1**	**56.2**	**36.9**	**16.2**	**35.8**	**46.6**	**37.5**	**26.2**	**34.8**

Table 1: Performance comparison of our proposed method with existing Video LLMs on three Video Grounding datasets (ActivityNet, Charades-STA, and DiDeMo). Methods are grouped into 7B and 13B series. The best results within each model series are shown in bold.

3.3 Data Integration

After constructing the Boundary Perception and Temporal Reasoning tasks, we integrate them into our training pipeline. TPE-VLLM is straightforward:

1) **Task Combination**: We uniformly mix the newly generated tasks with the original video grounding data. This ensures that the model is exposed to both the primary video grounding task and the new temporal reasoning tasks during training.

2) **Data Volume**: We generate approximately 20,000 new task instances in total. This relatively small addition (compared to 128,000 entries from stage 2 and 558,000 entries from stage 1) is designed to enhance the model's temporal reasoning abilities without overwhelming the core video grounding objective.

3) **Integration**: These new tasks are directly incorporated into the third stage of training, alongside the original video grounding data.

This simple yet effective integration approach allows our model to benefit from the additional temporal reasoning tasks while maintaining its focus on the primary video grounding objective.

4 Experiments

4.1 Dataset

To evaluate the effectiveness of our model, we conducted experiments on three publicly available datasets: ActivityNet Captions (Krishna et al., 2017), DiDeMo (Anne Hendricks et al., 2017), and Charades-STA (Gao et al., 2017).

ActivityNet Captions. This dataset comprises 10,009/ 4,917/ 5,044 videos in the training, validation, and test sets respectively, with corresponding query counts of 37,417/ 17,505/ 17,031. The average video duration is 117.6 seconds, and the average query length is 37.14 words. We adhere to the original dataset's splitting strategy and will report our results on the val_2 split.

Charades-STA. Charades-STA encompasses complex human behaviors and activities. The training set consists of 5,338 videos with 12,408 queries, while the test set contains 1,334 videos with 3,720 queries. The average video duration is 30.06 seconds, and the average query length is 7.22 words. We will report our results on the test set. It is noteworthy that we will utilize Charades-STA to evaluate our model's out-of-distribution generalization capability. The training data does not include any training data from the Charades-STA dataset.

DiDeMo. This dataset is divided into 8,395/ 1,065/ 1,004 videos for the training, validation, and test sets respectively, with corresponding query counts of 33,580/ 4,260/ 4,016. Each video is approximately 30 seconds in length, with an average query length of 7.5 words. We adhere to the original dataset's splitting strategy and will report our results on the test set.

4.2 Metric

We employ the recall rate "$R@n, IoU = m$", and $mIoU$ to evaluate our model's performance. Where R represents the proportion of queries where at least one of the k predictions has an IoU greater than m with the ground truth. $mIoU$

Boundary Perception		Temporal Understanding				ActivityNet				Charades-STA			
Duration	Position	ETM	EO	ES	TS	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
✘	✓	✓	✓	✓	✓	46.84	33.19	17.64	33.58	53.64	32.77	14.41	33.58
✓	✘	✓	✓	✓	✓	47.01	33.40	17.83	33.87	53.93	31.22	13.58	33.26
✓	✓	✘	✓	✓	✓	48.63	34.04	18.60	35.02	53.70	31.35	13.65	33.09
✓	✓	✓	✘	✓	✓	49.50	34.67	18.76	35.53	54.27	31.95	13.36	33.54
✓	✓	✓	✓	✘	✓	48.60	34.06	18.72	34.98	55.13	32.87	14.28	34.38
✓	✓	✓	✓	✓	✘	47.03	32.32	17.57	33.88	53.83	33.91	16.24	34.46
✘	✘	✓	✓	✓	✓	43.55	30.91	16.82	31.74	53.07	32.68	15.29	33.63
✓	✓	✘	✘	✘	✘	44.44	30.58	16.64	32.37	52.37	32.05	14.91	33.09
✓	✓	✓	✓	✓	✓	50.39	35.35	19.24	36.13	55.26	33.15	13.57	34.49

Table 2: Ablation studies on Boundary Perception and Temporal Understanding tasks for video grounding on TPE-VLLM(7B). Performance is evaluated on ActivityNet and Charades-STA datasets. Checkmarks (✓) indicate remaining tasks, while red crosses (✘) denote removed relevant task.

represents the mean Intersection over Union between the predicted intervals and the ground truth across all test samples. Consistent with previous works (Huang et al., 2024), we set $k = 1$ and $m = \{0.3, 0.5, 0.7\}$ for evaluation.

4.3 Implementation Details

For the visual encoder, we use CLIP ViT-L/14 (Radford et al., 2021) to extract visual features and keep the encoder frozen. For the Adapter, we use a linear layer and train it in the stage 1. For the LLM, we use Vicuna v1.5 (Chiang et al., 2023) for 7B and 13B, and fine-tuning them with LoRA, the parameters set to $r = 64$ and $alpha = 128$. We use the AdamW (Loshchilov, 2017) optimizer and set the learning rate to 0.0001. Our 7B model is trained on 1 RTX4090 GPU, while the 13B model is trained on 4 NVIDIA A100 GPUs.

4.4 Baselines

We compared our model with existing SOTA video-LLMs. These include VideoChat (Li et al., 2023), VideoLLaMA (Zhang et al., 2023), and VideoChatGPT (Maaz et al., 2023), which are designed for video understanding and interaction that have been fine-tuned on large-scale video-text pairs. There are also GroundingGPT (Li et al., 2024), VTimeLLM (Huang et al., 2024) and TimeChat (Ren et al., 2024), which are video large models specifically designed for fine-grained video temporal tasks.

4.5 Main Results

As shown in Table 1, the experimental results on three widely used video grounding datasets highlight the superiority of our model across different parameter settings. We compare our proposed approach to SOTA Video-LLMs, leading to several key findings, which underscore the potential of our approach in bridging the gap between language models and fine-grained video understanding.

For 7B series, TPE-VLLM consistently surpasses existing 7B Video LLMs across all datasets and evaluation metrics. On ActivityNet, it achieves notable gains compared to the next best model, VTimeLLM (7B parameters), with relative improvements of 14.5%, 27.3%, 34.3%, and 18.8% for R@0.3, R@0.5, R@0.7 and *mIoU*, respectively. On DiDeMo, TPE-VLLM also posted relative gains of 1.7%, 3.5%, 7.7% and 3.9% on R@0.3, R@0.5, R@0.7 and *mIoU*, respectively. On Charades-STA, we tested the out-of-distribution performance, compared to our baseline VTimeLLM, our results achieved relative gains of 8.8%, 20.3%, 29.0% and 11.2% in R@3, R@5, R@7 and *mIoU*, respectively. Concurrently, when compared to the current state-of-the-art model on this dataset, Time Chat, our model still achieved relative improvements of 2.8% and 9.7% in R@0.5 and R@0.7, respectively. The ability of our model to maintain high performance in the face of entirely new and unseen data distributions demonstrates the sophistication and robustness of our model. While scaling to larger model sizes, TPE-VLLM obtain notable performance improvements observed across all datasets. The experiments demonstrate strong generalization and consistent performance across multiple datasets with varying distributions. Specifically, on ActivityNet, we achieved an average improvement of 7.58%, significantly surpassing the VTimeLLM-13B. These results highlight the effectiveness of our method in improving both boundary perception and fine-grained temporal reasoning.

Data Size	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
20%	53.51	32.55	14.72	33.59
40%	53.67	31.32	13.84	33.01
60%	54.05	31.13	13.33	33.22
80%	54.81	31.00	13.55	33.76
100%	55.26	33.15	13.57	34.49

Table 3: Performance comparison with different data sizes on Charades-STA dataset.

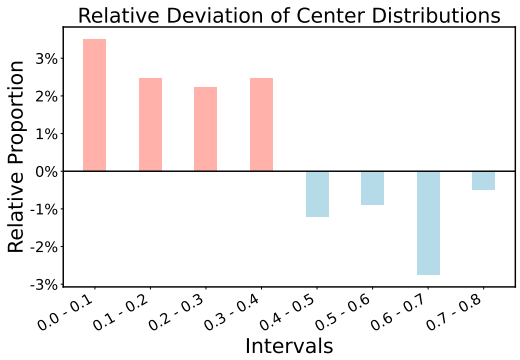


Figure 4: The relative deviation of the center point of our predicated results compared with baseline VTimeLLM. The center point deviation represents the degree of deviation between the center point of the forecast interval and the true interval. The lower the deviation, the more accurate the prediction of the event center.

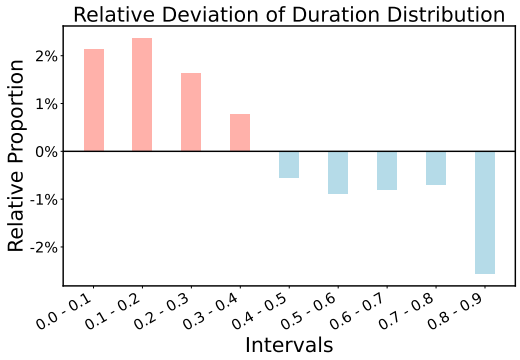


Figure 5: The illustration of the time deviation of our predicted results compared with the deviation of baselin VTimeLLM. The time deviation represents the difference between the duration of the predicted interval of the event and the duration of the true interval. The lower the difference, the closer the predicate time is to the true time.

4.6 Ablation Study

Based on the ablation studies depicted in Table 2, our experimental design methodically assesses the impact of removing specific tasks related to boundary perception and temporal understanding in video grounding on the TPE-VLLM (7B) model. These studies are executed across the ActivityNet and Charades-STA datasets.

The ablation study clearly shows that removing all tasks related to Boundary Perception and Temporal Understanding significantly influences the model's performance. Specifically, tasks associated with boundary perception are particularly impactful due to their strong alignment with the evaluation metrics. This alignment not only validates our model's effectiveness but also underscores the crucial role of boundary tasks in achieving superior performance. Removing any individual granular tasks also impacts the model's performance, thereby affirming the significance of each component. This outcome indicates that the tasks we have designed are complementary, they work synergistically to mitigate the impact of removing other tasks, thus enhancing the model's resilience. While the integration of all tasks yields the best results, the influence on the Charades dataset is notably milder compared to ActivityNet. This discrepancy may stem from the shorter video lengths and larger event proportions in Charades, which could diminish the impact of boundary ambiguities on performance metrics. Further exploration of these effects is discussed in subsequent sections 4.8.

4.7 Error Analysis

4.8 Data Size Impact Analysis

The original training dataset comprised of 16,128 entries. After partitioning 30% for validation purposes, we utilized 3720 entries for experiments of TPE-VLLM. To elucidate the impact of our methodology on the model's temporal understanding, rather than merely the increase in data volume, we conducted experiments with 20%, 40%, 60%, and 80% of the total constructed dataset. As illustrated in Table 3, employing only 20% of the dataset achieves results comparable to those obtained using the full dataset. This finding underscores the efficiency of our approach in leveraging smaller data volumes effectively. As the data size increases to 40%, 60%, 80%, and eventually 100%, there is a noticeable improvement in performance across all metrics, which highlights the capability of our method to bridge the gap between video temporal perception and textual temporal understanding. This incrementally enhances the model's performance, demonstrating the mitigation of impacts due to continuity and information density in the dataset. This nuanced approach effectively lever-

ages increasing amounts of data to progressively refine the model's temporal understanding capabilities, validating the effectiveness of our method in enhancing temporal alignment between the visual and textual modalities.

We compared the differences in the proportions of data within various fine-grained error intervals between our method and the SOTA model VTimeLLM.

We report the relative deviations in both the central point prediction and the duration prediction between our model and VTimeLLM, which is the difference in deviation between our model's predictions and those of VTimeLLM. Positive values indicate that TPE-VLLM has a higher proportion in these error intervals, while negative values indicate that the VTimeLM is higher. The central point value represents the temporal center at which the event occurred. A lower range of deviation indicates that the predicted center is closer to the actual center. A higher quantity of data within the lower deviation range, and a lower quantity within the higher deviation range, signifies that a greater number of cases have predicted the true event center more accurately, while fewer cases have predicted the incorrect event center. This indicates a better predictive performance. Similarly, the duration deviation represents the accuracy of the predicted event duration. The greater the amount of data with smaller deviations, the better the performance of the model.

Figure 4 shows the deviation of the predicted interval center point from the true center across different intervals. Notably, our model exhibits a higher proportion of predictions within smaller error ranges and significantly fewer predictions in larger error ranges compared to the VTimeLM. This indicates that our model achieves a more precise center point prediction, effectively reducing the instances of large errors. Figure 5 displays the relative lengths of the predicted intervals compared to the actual intervals. TPE-VLLM demonstrates a greater proportion of predictions within a tight error margin and fewer predictions with substantial deviation relative to the VTimeLM. This pattern suggests that our method not only predicts more accurate interval lengths but also maintains consistency in predicting closer to the true interval size, even under varying conditions.

5 Conclusion

In this paper, we introduced Time-Perception Enhanced Video Grounding via Boundary Perception and Temporal Reasoning, a method designed to improve large language models' temporal awareness in video grounding tasks. To address the limitations in LLMs' temporal understanding of video data, we designed two categories of specialized tasks: Boundary Perception Tasks and Temporal Reasoning Tasks. These tasks enable LLMs to more accurately identify event boundaries and understand temporal relationships between events in videos. By incorporating such diverse boundary perception and temporal reasoning tasks, TPE-VLLM achieved competitive performance on three public datasets and demonstrated the effectiveness of temporally-focused task design in enhancing LLMs' fine-grained video understanding capabilities.

6 Limitations

A limitation of our current approach is its reliance on labeled dataset construction. The construction of temporal reasoning and boundary perception tasks depends heavily on original labeled video grounding datasets. Future work could focus on exploring self-pity and reasoning to alleviate this dependence on labeled data. Such advancements may help improve the model's ability to generalize to plain video datasets and further enhance its reasoning capabilities, allowing it to infer more complex temporal dynamics.

Acknowledgements

We thank the COLING-2025 anonymous reviewers for their valuable feedback. This work was supported by the National Natural Science Foundation of China, 62106010 and the Technical Field Foundation (No. 2023-JCJQ-JJ-0747).

References

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

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Meng Cao, Fangyun Wei, Can Xu, Xiubo Geng, Long Chen, Can Zhang, Yuexian Zou, Tao Shen, and Daxin Jiang. 2023. Iterative proposal refinement for weakly-supervised video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6524–6534.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

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*, pages 5267–5275.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. 2023. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. 2024. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6657–6678.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. 2024. Zero-shot video moment retrieval from frozen vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5464–5473.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

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

Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Parallel attention network with sequence matching for video grounding. *arXiv preprint arXiv:2105.08481*.

Minghang Zheng, Shaogang Gong, Hailin Jin, Yuxin Peng, and Yang Liu. 2023. Generating structured pseudo labels for noise-resistant zero-shot video sentence localization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14197–14209.