

DisTime: Distribution-based Time Representation for Video Large Language Models

Yingsen Zeng^{1*}, Zepeng Huang^{1*}, Yujie Zhong^{1†}, Chengjian Feng¹, Jie Hu¹,
 Lin Ma¹, Yang Liu²

¹ Meituan Inc.

² Wangxuan Institute of Computer Technology, Peking University

{yingsen_2, huang-zpeng}@163.com jaszhong@hotmail.com yangliu@pku.edu.cn

Abstract

Despite advances in general video understanding, Video Large Language Models (Video-LLMs) face challenges in precise temporal localization due to discrete time representations and limited temporally aware datasets. Existing methods for temporal expression either conflate time with text-based numerical values, add a series of dedicated temporal tokens, or regress time using specialized temporal grounding heads. To address these issues, we introduce **DisTime**, a lightweight framework designed to enhance temporal comprehension in Video-LLMs. **DisTime** employs a learnable token to create a continuous temporal embedding space and incorporates a **Distribution-based Time Decoder** that generates temporal probability distributions, effectively mitigating boundary ambiguities and maintaining temporal continuity. Additionally, the **Distribution-based Time Encoder** re-encodes timestamps to provide time markers for Video-LLMs. To overcome temporal granularity limitations in existing datasets, we propose an automated annotation paradigm that combines the captioning capabilities of Video-LLMs with the localization expertise of dedicated temporal models. This leads to the creation of **InternVid-TG**, a substantial dataset with 1.25M temporally grounded events across 179k videos, surpassing ActivityNet-Caption by 55 times. Extensive experiments demonstrate that **DisTime** achieves state-of-the-art performance across benchmarks in three time-sensitive tasks while maintaining competitive performance in Video QA tasks. **DisTime** is released at <https://github.com/josephzpng/DisTime>.

1. Introduction

With Large Language Models (LLMs) displaying powerful comprehension abilities, significant efforts have been made to bridge the gap between language and vision, en-

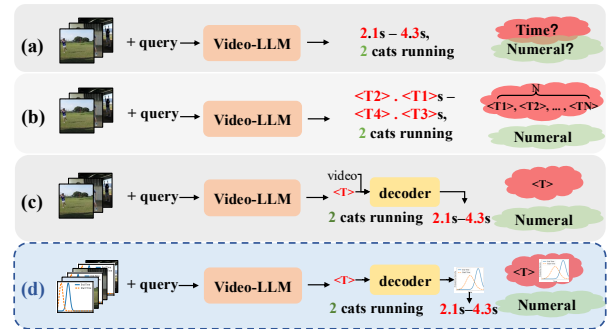


Figure 1. Comparison of temporal expression paradigms in Video-LLMs: (a) Text-Modal Discretization: time represented via text/number hybrids, (b) Multi-Token Discretization: dedicated temporal tokens, (c) Specialized Temporal Heads: utilizing temporal grounding models for direct regression, and (d) **DisTime**: distribution-based time decoder for temporal localization.

abling LLMs to tackle visual tasks. As a result, Image-LLMs [6, 7, 19, 27, 34, 57] have achieved breakthroughs in image-related tasks. In the realm of video tasks, which contain richer information, numerous Video-LLMs [6, 18, 20, 47, 49, 53] have also made substantial progress in areas such as reasoning, captioning, and summarization. However, for temporal comprehension, these models typically manage temporal information by capturing the sequence of events rather than the precise timing of their occurrence. This leads to significant shortcomings in time-sensitive tasks, such as moment retrieval [9, 12, 16, 25, 39–41], dense video captioning [16, 56], and grounded video question answering [3, 46], etc.

Temporal expression poses a significant challenge in time-sensitive Video-LLMs. The simplest approach is to express time using numbers, either alone or with markers, within the text modality. Models like GroundingGPT [24], VTimeLLM [15], and TimeMarker [4] employ this strategy. However, these methods force time and numbers to share the same decision boundary, complicating the classification process, as illustrated in Fig. 1a. To address modal

* Equal contribution with random order. † Corresponding author.

confusion, some methods like Momentor [35] and VTG-LLM [13] introduce extra tokens specifically for time expression, as shown in Fig. 1b. This approach, however, heavily depends on the classification of these new tokens and the balanced distribution of time data during training. Some tokens may receive insufficient training due to long-tailed distributions in the dataset. Both approaches share the drawback that using tokens for time expression is inherently discrete, resulting in insufficient precision for time representation (e.g., expressing decimals) and lacking explicit modeling of relationships between adjacent numerical values. The third approach, exemplified by InternVideo2.5 [44] and shown in Fig. 1c, creates trainable task tokens and a specialized temporal perception head to assist in time point decoding. These modules (e.g. CG-DETR [33]) are typically designed for temporal grounding, containing numerous parameters and requiring the re-input of visual information.

To improve the representation of temporal positions in LLMs, we propose **DisTime**, a method using an additional learnable token to indicate event time spans. This token creates a continuous embedding space for timestamps, thus avoiding confusion with numerical text values. As shown in Fig. 1d, DisTime utilizes the **Distribution-based Time Decoder** to convert the time token into start and end timestamps. Inspired by DFL [22], which acknowledges the inherent ambiguity in boundary detection, temporal boundaries face similar challenges, making direct regression to absolute time values difficult. Decoding the time token into a distribution provides flexibility when event boundaries are unclear, thereby reducing complexity in model prediction. Our method first converts the time token into a probability distribution, and then integrates this distribution to produce specific values. Unlike the approach in Fig. 1c, our decoder utilizes a minimal number of parameters (almost negligible) and without re-inputting original/encoded images.

Another challenge for time-sensitive Video-LLMs is the *scale of temporal-aware datasets*. Current video training datasets [2, 21, 51] predominantly focus on captioning, summarizing, and reasoning, often neglecting temporal awareness. The labor costs of annotating large volumes of time-sensitive data are significant. Although methods like VTimeLLM [15], InternVid-MR [11], and Momentor [35] propose unsupervised approaches to expand temporal grounding data, they share a common limitation: *temporal granularity constraint*, referring to reliance on shot boundaries and coarse fixed temporal intervals. VTimeLLM uses detected shot boundaries as time points, making event localization prone to errors within single shots, as simple boundaries do not accurately capture event timing. InternVid-MR divides videos into 2-second segments, determining event boundaries based on visual-text similarity, which often results in low labeling accuracy, par-

ticularly for short-duration events. Momentor merges segments by assessing the consistency of character instances between consecutive shots, but it still depends on the temporal boundaries of shots, ignoring static objects like scenery.

To expand temporal-aware video data, we propose an annotation paradigm that overcomes temporal granularity constraints for event boundaries. Our method harnesses the captioning abilities of Video-LLMs for event extraction and employs the fine-grained localization capabilities of dedicated models for event boundary detection. Drawing on expertise from well-annotated data, the pseudo labels break free from shot-boundary constraints, making them more suitable for temporal grounding tasks. We employ three dedicated models: UniMD [52] for traditional in-domain temporal grounding, Mr.Blip [32] for high-precision localization with LLMs, and TFFVTG [55] for zero-shot settings without training. A scoring strategy evaluates the grounding results, selecting the highest-scoring model output as the ensemble result for each event. Ultimately, we generated 1.25 million event annotations, over 55 times larger than the ActivityNet-Caption [16] dataset.

To summarize, this paper makes three contributions:

- We propose DisTime, a method that uses a single token for temporal expression in Video-LLMs. This approach transforms the token into a distribution via a distribution-based time decoder, which addresses event boundary uncertainty and enables continuous time modeling.
- We propose an automated annotation paradigm that combines the captioning capabilities of LLMs with the fine-grained grounding abilities of dedicated temporal grounding models. With this paradigm, we create the **InternVid-TG** dataset, featuring 179K videos and 1.25M events.
- Experiments demonstrate that DisTime enhances the fine-grained grounding abilities of Video-LLMs across three time-sensitive tasks. Notably, on Charades-STA [9], our model surpasses all dedicated models and current Video-LLMs in $R@1_{iou=0.3}$ score, even in a zero-shot setting.

2. Related Work

Time-sensitive video understanding. Time-sensitive video understanding tasks require interpreting a video as a sequence of interconnected events and locating them in a temporally grounded manner. Moment Retrieval [9, 12, 16] focuses on identifying the timing of a specific event within a video. Dense Video Captioning [16, 56] involves describing and summarizing all events with their temporal information along the video timeline. Grounded Video Question Answering [3, 46] demands understanding questions while addressing the embedded temporal information. Thus, these tasks challenge models to have strong temporal sensitivity.

Vision language models. With advancements in Large Language Models (LLMs), there is a growing focus on extending their capabilities from natural language process-

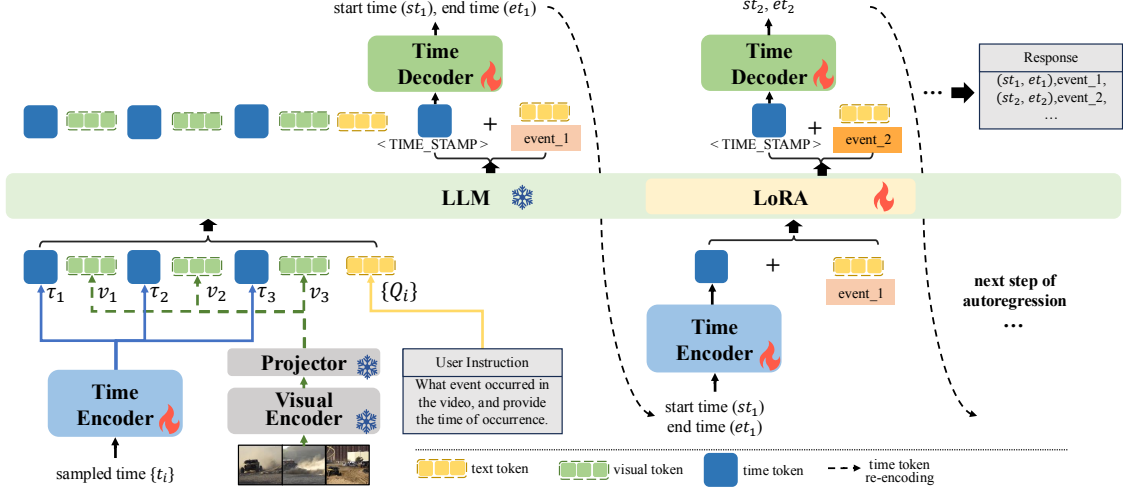


Figure 2. **Overview of the Video-LLM with DisTime.** Our DisTime method employs a single token (denoted as $\langle \text{TIME_STAMP} \rangle$) to represent continuous time, transformed by a distribution-based time decoder and time encoder. Initially, sampled timestamps are converted into time tokens and integrated into the LLM’s input sequence. During autoregressive generation, when a $\langle \text{TIME_STAMP} \rangle$ is encountered, its corresponding LLM last-layer embedding is decoded into start and end times. These timestamps are then re-encoded to update the temporal context for subsequent autoregressive steps. In the final output sequence, the decoded times replace the $\langle \text{TIME_STAMP} \rangle$. This approach enables the Video-LLM to explicitly model time, thereby enhancing its time-sensitive capabilities.

ing to the visual domain, enhancing functions such as question answering, description, and summarization based on visual information. Recent Vision Language Models [6, 7, 19, 27, 57] incorporate learnable modules to integrate language and vision, primarily targeting image-level perception and comprehension. As multimodal research has progressed, Video-LLMs [18, 21, 23, 53] have emerged, showing strong performance in tasks like video question answering and captioning. However, these approaches largely concentrate on visual-level task requirements and exhibit limitations in high-level temporal perception.

Temporal-awareness video language models. To tackle the challenges of time-sensitive tasks, recent Video-LLMs have focused on constructing time-sensitive data and enhancing temporal awareness in their architectures. GroundingGPT [24] improves performance by generating time-sensitive datasets and integrating multimodal data. VtimeLLM [15] enhances temporal perception through optimized training strategies, enriched datasets, and tailored instruction. TimeChat [37] and TimeMarker [4] incorporate temporal information from the current video frame into the multimodal input, enabling systems to perceive and interpret temporal dynamics effectively. However, these methods model time using textual representation, sharing decision boundaries with regular numerical values, thereby complicating classification and introducing quantization errors. Momentor [35] and VTG-LLM [13] introduce special tokens for timestamp modeling. Momentor uses a temporal perception module to ensure time continuity, while VTG-LLM incorporates absolute-time tokens to manage timestamp information, reducing quantization errors. Despite

this, these methods rely heavily on numerous new tokens for representing time absolutely, often overlooking time continuity and boundary fuzziness. We aim to maintain time continuity while addressing boundary fuzziness attentively.

3. Method

This section delves into the proposed DisTime, a method that uses a single token to represent continuous time in large language models (LLMs), enabling transformation between time tokens and timestamps via **Distribution-based Time Encoder-Decoder**. In Sec. 3.1, we present an overview of our approach to time-sensitive video understanding. Sec. 3.2 details the rationale and strategy for modeling time as a distribution in LLMs. We then introduce the Iterative Time Refinement mechanisms (Sec. 3.3) that update the temporal context during the autoregression process. Finally, Sec. 3.4 outlines the objective function and training methodology for optimizing the time token, time decoder, and time encoder.

3.1. Overview

The architecture of our time-sensitive video-LLM with DisTime is illustrated in Fig. 2, comprising five core components: a visual encoder with a projector, a text encoder, a LLM, a time decoder ($\Phi_{\text{time-dec}}$) and a time encoder ($\Phi_{\text{time-enc}}$). First, a certain number of video frames are uniformly sampled from the video. Then, the visual encoder encodes the video frames, and the projector maps them into the language space to obtain visual tokens $\{V_i\}_{i=1}^T$, where T is the number of sampled time steps in the video. Then, we derive a sequence of temporal representations $\{\tau_i\}_{i=1}^T$

by processing the timestamps through the time encoder, as:

$$\tau_i = \Phi_{\text{time-enc}}(t_i), i = 1, 2, \dots, T \quad (1)$$

where $t_i \in [0, 1]$ is the relative timestamp normalized to the video duration. Subsequently, the time tokens $\{\tau_i\}_{i=1}^T$ and visual tokens $\{v_i\}_{i=1}^T$ are interleaved and jointly input to the LLM with the user's instruction text tokens $\{Q_i\}_{i=1}^N$, as:

$$\text{Input} = \text{Concat}(\tau_1, v_1, \dots, \tau_T, v_T, Q_1, \dots, Q_N), \quad (2)$$

where $\{Q_i\}$ is encoded via text encoder. We use a dedicated time token, termed **Distribution-based Time Token**, to represent continuous time. This token is instantiated in the vocabulary as `<TIME_STAMP>`. The autoregressive generation process includes both standard vocabulary tokens and the dedicated time token. When the LLM outputs `<TIME_STAMP>`, we extract the corresponding hidden state $\mathbf{h}_{\text{time}}$ from the final layer and pass it to the time decoder to obtain continuous timestamps, as follows:

$$st, et = \Phi_{\text{time-dec}}(\mathbf{h}_{\text{time}}), \quad (3)$$

where $st \in [0, 1]$ and $et \in [0, 1]$ denote the decoded start time and end time. These timestamps are then reprocessed through the time encoder to refresh the temporal context for subsequent autoregressive steps. The `<TIME_STAMP>` is replaced with the re-encoded time token, establishing explicit temporal referencing that functions as video timeline pointers. Finally, the `<TIME_STAMP>` is replaced with corresponding decoded timestamps (st, et), which are concatenated with generated text tokens (e.g. `event_2`) to form the LLM's complete output. This produces time-sensitive predictions with explicit timestamps.

Our method is compatible with mainstream Video-LLMs (e.g. InternVL2.5 [5], LLaVA-OneVision [18]) that retain temporal-aligned input token sequences while being incompatible with the models using global temporal aggregation (e.g. LinVT [10], BLIP3-Video [38]).

3.2. Modeling Time as Distribution

DisTime to LLMs. Current Video-LLMs excel in global video understanding but often struggle with time-sensitive tasks due to imprecise event timing modeling. Existing solutions either (1) express time in text modality (complicating the classification for time and numbers), (2) use numerous learnable tokens (suffering from long-tailed distributions and temporal discontinuity), or (3) employ heavy temporal modules (adding extra high computational costs). The representation of time remains a critical optimization target for time-sensitive modeling.

Furthermore, inspired by [22], we recognize the inherent ambiguity in event boundary detection. For instance, the query ‘‘A person is drinking water from a cup’’ introduces annotation uncertainty, such as whether to include the cup-picking action in the temporal segment. This ambiguity makes direct regression to absolute timestamps susceptible

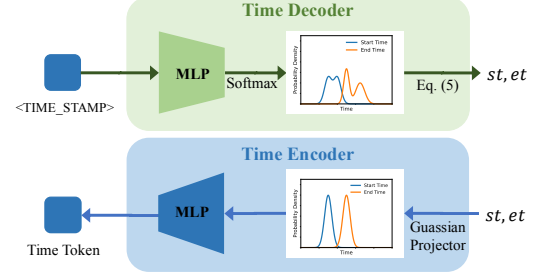


Figure 3. The illustration of time decoder and time encoder.

to precision errors. Holistic analysis of multiple boundary hypotheses is essential to mitigate uncertainty.

To address these limitations, we propose DisTime - a time modeling paradigm that represents time as probability distributions within LLMs. Our framework achieves continuous temporal representation through two key innovations: (1) Single-Token Constraint: This encapsulates continuous temporal information within a dedicated token using a lightweight encoder ($\Phi_{\text{time-enc}}$) and decoder ($\Phi_{\text{time-dec}}$), enabling precise temporal grounding with minimal computational overhead. (2) Distribution-Aware Decoding: This resolves boundary ambiguity through probabilistic timestamp generation, where final temporal coordinates are derived from weighted combinations of multiple hypotheses.

Distribution-based time token. The dedicated time token in LLMs is termed Distribution-based Time Token (denoted as `<TIME_STAMP>`), distinctly separating it from numerical text tokens in the vocabulary. When the LLM generates a `<TIME_STAMP>`, a lightweight decoder with a softmax function transforms it into a latent probability distribution, with each dimension of the latent embedding corresponding to an anchor point on the normalized time axis. These probability distributions enable continuous value representation.

Time decoder. The time decoder is used to decode the `<TIME_STAMP>` into continuous timestamps, as shown in Fig. 3. Specifically, we partition the normalized temporal axis into $reg_{max} + 1$ discrete bins, each linked to an anchor point representing the probability of the timestamp being localized at that position. Formally, the temporal distribution is represented as a vector $\mathbf{e} = [\mathbf{e}_{st}, \mathbf{e}_{et}]$, where $\mathbf{e} \in \mathbb{R}^{2 \times (reg_{max} + 1)}$, and $\mathbf{e}_{st}$ and $\mathbf{e}_{et}$ denote the start and end time distributions, respectively. A lightweight Multi-Layer Perceptron (MLP) followed by a softmax activation maps the `<TIME_STAMP>` to distribution vectors $\mathbf{e}$, as:

$$\mathbf{e} = \Phi_{\text{time-dec}}(\mathbf{h}_{\text{time}}) = \text{softmax}(\text{MLP}(\mathbf{h}_{\text{time}})). \quad (4)$$

Continuous timestamps of start time and end time are derived via anchor-weighted summation, as:

$$st = \sum_{i=0}^{reg_{max}} \mathbf{e}_{st}^{(i)} \cdot a_i, \quad et = \sum_{i=0}^{reg_{max}} \mathbf{e}_{et}^{(i)} \cdot a_i, \quad (5)$$

where $\{a_i\}_{i=0}^{reg_{max}}$ are anchors defined as: $a_i = i/reg_{max}$.

Time encoder. As shown in Fig. 3, the time encoder acts as the inverse of the decoder, translating continuous timestamps $(st, et) \in [0, 1]$ into a time token τ . Firstly, each

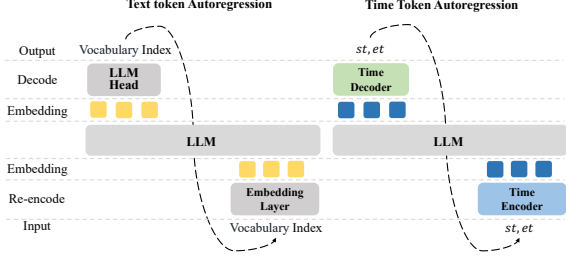


Figure 4. The autoregression of text token and time token.

timestamp is projected into a Gaussian-regularized distribution to model annotation ambiguity, as: $p_{st} \sim \mathcal{N}(st, \delta^2)$, $p_{et} \sim \mathcal{N}(et, \delta^2)$, where δ controls the distribution spread. Here $\delta = 1$ for boundary uncertainty. The continuous distributions p_{st} and p_{et} are then discretized into $reg_{max} + 1$ uniformly spaced bins spanning $[0, 1]$ to obtain the discretized distributions $\hat{e}_{st}$ and $\hat{e}_{et}$. Next, a MLP projects the concatenated distributions $[\hat{e}_{st}, \hat{e}_{et}]$ into LLM-compatible time tokens τ , as:

$$\tau = \text{MLP}([\hat{e}_{st}, \hat{e}_{et}]). \quad (6)$$

The encoding process operates through two stages: (1) time token injection during input visual sequence composition, and (2) LLM autoregressive processing. In the former stage, we set $st = et$ to model instantaneous temporal positions. In the latter stage, we utilize timestamps decoded by the time decoder and replace the `<TIME_STAMP>` with the encoded token τ for refined temporal representation.

3.3. Iterating Time Refinement

Time token re-encoding. The hidden embedding of `<TIME_STAMP>` inherently capture event boundary ambiguity (e.g. multi-modal distributions). In the autoregressive generation paradigm of LLMs, retaining raw time tokens across iterations can propagate errors due to representation uncertainty. To address this, we use probability-weighted timestamps from the decoder output to refine the LLM’s temporal context. As shown in Fig. 3, we re-encode the decoded timestamps into fixed-variance Gaussian distributions via the time encoder. This re-encoded time token serves dual purposes: it transforms ambiguous temporal token interpretations into well-defined Gaussian-distributed representations and ensures distributional alignment across all time tokens. Through this method, arbitrary decoded distributions are normalized into standardized Gaussian forms, enabling consistent temporal representation and enhancing the LLM’s temporal comprehension capability.

Time token autoregression. The time token autoregression operates in a similar fashion to text tokens, as depicted in Fig. 4. Both token types undergo explicit decoding: text tokens map to vocabulary indices, while time tokens decode to continuous timestamps (st, et). The decoded results are then transformed into embeddings for subsequent iterations via domain-specific encoding: an embedding layer for text tokens and a time encoder for time tokens.

3.4. Training Strategy

Training Object. The loss function comprises three components: distribution focal loss $\mathcal{L}_{dist}$ for temporal distribution learning, 1d-IoU regression loss $\mathcal{L}_{reg}$ to directly optimize temporal grounding, and next token prediction loss $\mathcal{L}_{ntp}$ for autoregressive generation. The distribution focal loss [22] is key for training the distribution-based time token and time decoder, as it learns the underlying general distribution without introducing additional priors. The unified training objective is formulated as: $\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{ntp} + \lambda_2 \cdot \mathcal{L}_{reg} + \lambda_3 \cdot \mathcal{L}_{dist}$, where we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ for balanced multi-task optimization.

Training Recipe. Based on the Video-LLMs that have been trained for general understanding capabilities, our method trains only one stage for enhancing temporal grounding ability. During this stage, the vision backbone and intermediate layers are frozen, and LoRA [14] is employed for efficient fine-tuning of the LLM. Additionally, the parameters of the LLM token embedding, LLM head, and the time encoder and decoder are fully trained.

4. InternVid-TG

Existing large-scale video training datasets primarily emphasize captioning, summarization, and reasoning tasks while lacking temporal awareness. Furthermore, though some automated annotation methods [11, 15, 35] attempt to scale supervision, they remain constrained by shot boundaries or coarse fixed temporal intervals. To expand time-aware supervision and overcome these constraints, we propose a temporal-aware data expansion framework that synergistically combines the summarization capabilities of large multimodal models with the fine-grained localization capabilities of temporal dedicated models. Fig. 5 illustrates the proposed annotation paradigm, which comprises four steps: event capturing, event boundary detection, scoring and ensemble processes, and instruction writing. Ultimately, we automate the annotation of InternVid-FLT [42] video data, transitioning from the original video-text alignment data to temporal grounding data, named **InternVid-Temporal-Grounding (InternVid-TG)**.

4.1. Generation pipeline

Event capturing. Video-LLMs are trained on a large amount of video caption data, which endows them with powerful descriptive capabilities. We leverage this to extract event descriptions present in videos. First, we obtain the untrimmed sources from InternVid-FLT. We maintain the original resolution of the video and decode it at 1 fps to form an image sequence, which serves as the visual input for GPT-4o [1]. Next, we use the query “please identify the events that occur in the video” to prompt the model. In response, GPT-4o generates dense video captions in the

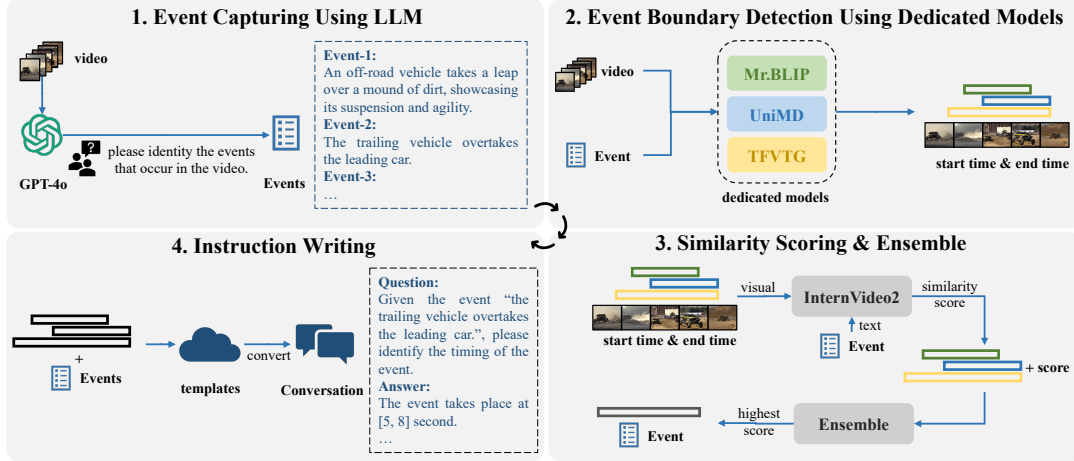


Figure 5. The pipeline of our temporal-aware data expansion paradigm, which utilizes multimodal LLM and dedicated temporal models to generate temporal grounding annotations.

format of “(start time, end time, event description)”. Here, we disregard the segment information in the response and use the event description as the extracted event. Finally, we obtain ≈ 7.0 events per video.

Event boundary detection using dedicated models. In the step of automatically annotating time points, to eliminate the dependency on shot segmentation, we directly use three models specifically designed for visual temporal grounding to perform the annotation. These models include UniMD [52], Mr.Blip [32], and TFVTG [55].

UniMD is a traditional method specialized for temporal grounding. We retrain it using the visual features from InternVideo2 [43] as input because the training data for the visual encoder of InternVideo2 includes videos from InternVid-FLT. This adaptation makes UniMD more suited to the domain of InternVid-FLT. Mr.Blip is a dedicated Video-LLM trained for temporal grounding, and it is one of the state-of-the-art (SOTA) models in the Charades-STA [9], ANet-Caption [16], and QVHighlight [17] benchmarks. In QVHighlight, it achieves a top-1 recall at 50% IoU ($R@1_{iou=0.5}$) of 74.77%, demonstrating a high degree of overlap in event localization. TFVTG is a train-free video grounding model adapted for zero-shot scenarios. In zero-shot situations, it achieves a top-1 recall at 30% IoU ($R@1_{iou=0.3}$) of 67.04% on the Charades-STA benchmark, showing competitiveness with the fine-tuned SOTA LD-DETR [54] which achieves 73.92% $R@1_{iou=0.3}$.

By inputting the extracted event descriptions and the videos, we independently input them into these three dedicated models to obtain the time segment for each event.

Similarity scoring and ensemble. To evaluate the segment output from step two, we introduce a similarity scoring strategy. This score subsequently serves as the basis for ensembling the segments from these three models. We utilize InternVideo2 [44] as our scoring model. Initially, each video segment is trimmed, and the video encoder of Intern-

Dataset	annotator	#videos	Avg Dura. (s)	#events	#events/vid
Ego4d [12]	Human	2k	472	18.4k	9.2
Charades-STA [9]	Human	10k	30	68k	6.8
ANet-caption [16]	Human	15k	118	22.5k	1.5
VTimeLLM [15]	LLM	146k	52	353k	2.4
InternVid-MR [11]	LLM	-	-	150k	-
Momentor [35]	LLM	65k	403	1460k	22.5
InternVid-TG (ours)	LLM+ dedicated	179k	55	1250k	7.0

Table 1. Comparison among InternVid-TG, existing human-annotated datasets, and unsupervised datasets.

Video2 is employed to extract visual features. Concurrently, the textual feature of each event caption is derived using InternVideo2’s text encoder. The cosine similarity between these features is computed to assess text-video relevance. For each event, the temporal result with the highest score is selected as the ensemble output. In the supplementary materials, we present evidence supporting the effectiveness of our scoring strategy and ensemble approach. Additionally, we use the human-annotated InternVid-TG-1k, comprising 1,000 queries, to assess the quality of the pseudo labels.

Instruction writing. We design 5 dialogue templates to convert temporal grounding annotations into single-turn conversations, which require identifying timestamps for the given event descriptions in the video.

4.2. Impact

Scale. Tab. 1 presents our dataset’s data distribution. InternVid-TG annotates 1.25M events from 179k videos, originating from the same source as VTimeLLM and InternVid-MR. However, we extract more events, resulting in the largest video count among these datasets. Overall, our method provides a scalable data expansion approach. Since InternVid-FLT includes 4 million videos, we can expand further to incorporate more videos if needed.

5. Experiment

In this section, we begin by detailing the experimental setup (Sec. 5.1) and benchmarks (Sec. 5.2). Following this, in

Dataset	Task
VideoChat2 [21] (232k)	General Video Understanding
ANet-Caption [16] (37k)	Moment Retrieval
ANet-Caption [16] (9.8k)	Dense Video Captioning
YouCook2 [56] (2k)	Dense Video Captioning
ET-Instuct [28] (126k)	Multi-Event
InternVid-TG (1250k)	Moment Retrieval

Table 2. Details of the training data covering both general video understanding and time-sensitive comprehension.

Sec. 5.3, we conduct ablation studies to examine the time distribution representation, time token re-encoding, and the effectiveness of our proposed dataset. Finally, in Sec. 5.4, we provide further comparisons with current Video-LLMs for time-sensitive and general QA tasks. Additional ablation studies, qualitative analyses, and further benchmarks are included in the supplementary material.

5.1. Implementation Details

The proposed DisTime is integrated into two current Video-LLMs, InternVL2.5 [5] and LLaVA-OneVision [18], to enhance its performance on time-sensitive tasks. We use InternVL2.5-1B, LLaVA-OneVision-7B, and InternVL2.5-8B as baselines, with ablation studies conducted on InternVL2.5-1B. For visual input, both Video-LLMs sample frames at equal intervals: InternVL2.5 samples 16 frames, while LLaVA-OneVision samples at most 32 frames. The MLP for the time encoder and decoder consists of three fully connected layers with ReLU functions. The time decoder and encoder are parameter-efficient, collectively comprising just 0.36% of InternVL2.5-1B, 0.34% of LLaVA-OneVision-7B and 0.84% of InternVL2.5-8B. The specific training data are detailed in Tab. 2.

5.2. Benchmarks

The comprehensive evaluation covers three time-sensitive tasks: Moment Retrieval (MR), Dense Video Captioning (DVC), and Grounded Video Question Answering (Grounded-VQA). For MR, we use the Charades-STA [9], ANet-Caption [16], and QVHighlights [17]. For DVC, we evaluate on ANet-Caption and YouCook2 [56]. For Grounded-VQA, we employ the NExT-GQA [46] dataset. Additionally, we assess our model’s performance on general video understanding tasks using MVBench [21], Video-MME [8], and LongVideoBench [45].

5.3. Ablation Study

Roles of distribution representation. The comparison between directly predicting timestamps (denoted as “Direct”) and obtaining timestamps through distribution prediction (denoted as “Dist.”) is presented in Tab. 3. Utilizing distribution representation for time enhances the model’s performance across all metrics in the MR task’s Charades-STA dataset and the DVC task’s YouCook2 dataset. Notably, YouCook2 exhibits a significant improvement, with the F1 score increasing from 2.2% to 16.3%.

Direct	Dist.	Re-Enc.	Charades-STA (MR)				YouCook2 (DVC)		
			R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU	SODA _c	CIDEr	F1 Score
✓			76.6	51.9	24.9	49.5	0.6	0.9	2.2
	✓		77.0	53.5	26.7	50.1	0.9	4.0	16.3
		✓	78.1	56.3	29.7	51.6	4.2	15.6	20.5

Table 3. Ablation study of different components in the time tokenizer. “Direct” indicates that the decoder directly predicts timestamps, while “Dist.” means the decoder first predicts a distribution and then aggregates it into timestamps. “Re-Enc.” denotes the use of time token re-encoding for the next step of autoregression.

Training Data	Charades-STA (MR)				QVHighlights (MR)			
	R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)	mAP@.5	mAP@.75
Baseline	77.4	51.4	22.9	48.7	38.7	17.7	34.8	12.5
w/ VTimeLLM [15]	76.2	51.9	25.0	49.3	51.0	25.8	46.0	18.8
w/ Momentor [35]	76.6	50.6	21.6	47.8	39.7	17.8	35.1	12.5
w/ InternVid-TG [†]	77.8	55.5	28.0	51.0	52.2	26.4	46.2	18.9
w/ InternVid-TG	78.1	56.3	29.7	51.6	54.1	27.8	47.9	19.2

Table 4. The effectiveness of the LLM-generating time-sensitive datasets. Baseline refers to the basic training data, which includes general video understanding and time-sensitive datasets excluding the datasets used for comparison. InternVid-TG[†] denotes a subset that includes videos overlapping with those in VTimeLLM.

Roles of time token re-encoding. As shown in Tab. 3, time token re-encoding (“Re-Enc”) enhances distribution representation effectiveness, particularly in the high-precision metrics required for the MR task on Charades-STA, where $R@1_{IoU=0.7}$ improves by 3%. This underscores the importance of time token re-encoding for high-precision temporal expressions. Furthermore, in the YouCook2 dataset for the DVC task, “Re-Enc” leads to an increase of 11.6% in the CIDEr score and a boost of 4.2% in the F1 score.

The effectiveness of InternVid-TG. We assess our proposed dataset’s effectiveness on two MR benchmarks: Charades-STA and QVHighlights. For a fair comparison with VTimeLLM, we used a subset (InternVid-TG[†]) that shares overlapping videos with VTimeLLM, totaling $\approx 90k$ videos. As Tab. 4 shown, compared to VTimeLLM and Momentor, the model with InternVid-TG[†] shows improvements on both benchmarks, even though Momentor contains more events than InternVid-TG[†]. Notably, annotation noise can decrease metrics, as seen with Momentor, where the model’s performance on Charades dropped. Finally, utilizing our full InternVid-TG dataset, expanding from 90k to 179k videos, further boosts performance.

5.4. Main Results

Moment Retrieval. As shown in Tab. 5, our model is compared to dedicated methods and time-sensitive Video-LLMs for the MR task on the Charades-STA, ANet-Caption benchmarks. Our DisTime-InternVL and DisTime-LLaVAOV models outperform all others in $R@1_{IoU=0.3}$, even in a zero-shot setting. Among Video-LLMs optimized for MR tasks (*i.e.* excluding InternVL2.5 and LLaVA-OneVision), DisTime demonstrates robust time-sensitivity. In Charades-STA, DisTime-InternVL-1B surpasses others across all metrics, notably outperforming TimeMarker by +4.4% in $R@1_{IoU=0.5}$. The DisTime-InternVL-8B model

Model	Size	Charades-STA				ANet-Caption			
		R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU	R@1 (IoU=0.3)	R@1 (IoU=0.5)	R@1 (IoU=0.7)	mIoU
Dedicated									
UniVTG [26]	-	72.6	60.2	38.6	52.2	-	-	-	-
UniMD [52]	-	-	63.9	42.2	-	-	42.7	25.1	-
Mr.BLIP [32]	3B	-	69.3	49.3	58.6	-	53.92	35.6	-
LLaVA-MR [30]	3B	-	70.7	49.6	59.8	-	-	-	-
Video-LLMs									
Momentor [35]	7B	42.6	26.6	11.6	28.5	42.9	23.0	12.4	29.3
ChatVTG [36]	7B	52.7	33.0	15.9	34.9	40.7	22.5	9.4	27.2
VTimeLLM [15]	13B	55.3	34.3	14.7	34.6	44.8	29.5	14.2	31.4
TimeChat [37]	7B	-	32.2	13.4	-	-	-	-	-
VTG-LLM [13]	7B	-	33.8	15.7	-	-	-	-	-
InternVideo2.5 [44]	7B	-	43.3	-	41.7	-	-	-	-
TimeMarker [4]	8B	73.5	51.9	26.9	48.4	67.4	50.7	33.0	49.5
InternVL2.5* [5]	1B	3.1	1.5	0.7	5.1	5.3	2.9	1.5	5.5
LLaVA-OneVision* [18]	7B	20.8	7.2	2.2	14.5	24.8	14.0	9.0	20.5
InternVL2.5* [5]	8B	27.9	10.9	4.2	18.2	21.1	12.7	8.6	17.2
DisTime-InternVL	1B	78.1	56.3	29.7	51.6	67.1	45.4	23.1	45.2
DisTime-LLaVAOV	7B	76.3	52.0	19.8	47.4	57.6	34.3	13.7	37.1
DisTime-InternVL	8B	81.0	60.3	30.8	53.1	72.9	53.2	30.3	50.5

Table 5. Comparison with various models on Moment Retrieval task in Charades-STA and ANet-Caption. “Dedicated” refers to the model fine-tuned with benchmark-specific training data. “*” indicates results from our supplementary testing of the model. Notably, in Charades-STA, our DisTime is in a zero-shot setting.

Model	Size	YouCook2			ANet-Caption		
		SODA.c	CIDEr	F1 Score	SODA.c	CIDEr	METEOR
Valley [31]	7B	0.1	0.0	1.5	0.3	1.8	0.8
Momentor [35]	7B	-	-	-	2.3	14.9	4.7
VTimeLLM [15]	13B	-	-	-	5.9	27.2	6.7
TimeChat [37]	7B	1.2	3.4	12.6	-	-	-
VTG-LLM [13]	7B	1.5	5.0	17.5	-	-	-
DisTime-InternVL	1B	4.2	15.6	20.5	4.5	17.9	4.7
DisTime-LLaVAOV	7B	5.4	21.8	19.5	4.9	21.3	4.9
DisTime-InternVL	8B	6.9	31.0	26.4	4.1	17.8	5.8

Table 6. Comparison with LLM-based models on the Dense Video Captioning task in ANet-Caption and YouCook2.

further widens this gap to +8.4%. In ANet-Caption, our 8B model excels in $R@1_{IoU=0.3}$ and mIoU, trailing slightly behind Mr.BLIP, which is dedicated to ANet-Caption, but remains competitive in $R@1_{IoU=0.5}$. Compared to the baseline InternVL2.5, DisTime significantly enhances the MR task performance. The 1B model demonstrates an improvement in $R@1_{IoU=0.3}$ on Charades-STA, increasing from 3.1% to 78.1%. This validates our method’s superior ability to express temporal information in LLMs and to enhance their skills in temporal localization. Furthermore, incorporating DisTime into LLaVA-OneVision considerably increases its temporal sensitivity, showcasing the generalizability and scalability of our method.

Dense Video Captioning. As shown in Tab. 6, in the step-wise dense video captioning task, both DisTime-InternVL and DisTime-LLaVAOV significantly outperform all Video-LLMs on the YouCook2 benchmark. This improvement is attributed to our proposed Iterative Time Refinement mechanism, which clarifies the output time token through re-encoding, thereby more accurately conveying the timing of the previous event in each subsequent autoregressive step. For the ANet-Caption benchmark, our method surpasses Valley and Momentor but falls short of VTimeLLM. It is important to note that, unlike VTimeLLM, which relies on 100-frame inputs, our method uses fewer input frames, with DisTime-InternVL using 16 frames and DisTime-LLaVAOV using 32 frames. Capturing nuanced event progression in ANet-Caption may require richer visual context from additional frames.

Grounded Video Question Answering. The results for the grounded video question answering task on the NExT-GQA

Model	Size	NExT-GQA							
		Acc@QA	Acc@GQA	mIoP	IoP@0.3	IoP@0.5	mIoU	IoU@0.3	IoU@0.5
SeViLA [30]	3B	68.1	16.6	25.7	34.7	22.9	21.7	29.2	13.8
Temp [46]	150M	60.2	16.0	25.7	31.4	25.5	12.1	17.5	8.9
FrozenBiLM [48]	1B	70.8	17.5	24.2	28.5	23.7	9.6	13.5	6.1
DisTime-InternVL	1B	65.9	17.0	33.5	42.9	23.7	31.1	41.0	21.8
DisTime-LLaVAOV	7B	75.8	22.8	36.3	48.4	28.7	31.9	45.6	23.5
DisTime-InternVL	8B	80.3	28.1	40.5	55.1	33.8	35.9	52.0	28.8

Table 7. Comparison with models on the Grounded Video Question Answering task in NExT-GQA.

Model	Size	MVBench avg	Video-MME (wo sub))	LongVideoBench (val total))
TimeChat	7B	38.5	34.7	-
TimeMarker	8B	67.4	57.3	56.3
InternVideo2.5	7B	75.7	65.1	60.6
InternVL2.5	1B	64.3	50.3	47.9
LLaVA-OneVision	7B	56.7	58.2	56.3
InternVL2.5	8B	72.0	64.2	60.0
DisTime-InternVL	1B	62.6	47.4	49.0
DisTime-LLaVAOV	7B	59.0	54.4	55.4
DisTime-InternVL	8B	72.0	59.6	60.0

Table 8. Comparison with models on the General Video Understanding tasks in short (MVBench and Video-MME) and long (LongVideoBench) video benchmarks.

benchmark are shown in Tab. 7. This task not only evaluates the understanding of question answering but also requires the ability to temporally locate events. Among the compared methods, DisTime-InternVL-1B surpasses them in most metrics (except for Acc@GQA and IoP@0.5). When scaled up to DisTime-InternVL-8B, it achieves the top performance across all metrics.

General Video Understanding. The results of the general video understanding task are shown in Tab. 8. Without any targeted optimization of general understanding capabilities, our methods exhibit diverse performance improvements across various benchmarks. In MVBench, DisTime-LLaVAOV-7B outperforms LLaVA-OneVision by 2.3%, while DisTime-InternVL-8B maintains consistent performance metrics. In LongVideoBench, DisTime-InternVL-1B shows a improvement of 1.1%. Despite a decline in performance on VideoMME, both DisTime-InternVL and DisTime-LLaVAOV remains competitive when compared to other models. For example, DisTime-InternVL-8B significantly outpaces TimeMarker with the same parameter size, and DisTime-InternVL-1B surpasses TimeChat, which has a much larger parameter size. The integration of DisTime primarily aims to enhance time-sensitive tasks, while still maintaining competitiveness in general understanding tasks.

6. Conclusion

This paper proposes DisTime, a lightweight paradigm for enhancing temporal expression in Video-LLMs by using a single token to regress continuous timestamps through distribution prediction. Additionally, we propose an automated annotation paradigm that leverages LLMs and specialized temporal grounding models. Video-LLMs equipped with DisTime achieve improved temporal comprehension through efficient continuous time modeling and scalable data generation, providing practical solutions for applications requiring fine-grained video understanding.

Acknowledgment

This work was supported by the grants from the National Natural Science Foundation of China 62372014 and Beijing Natural Science Foundation 4252040.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 2
- [2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 2
- [3] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024. 1, 2
- [4] Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024. 1, 3, 8
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4, 7, 8
- [6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1, 3
- [7] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 1, 3
- [8] Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 7
- [9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 1, 2, 6, 7
- [10] Lishuai Gao, Yujie Zhong, Yingsen Zeng, Haoxian Tan, Dengjie Li, and Zheng Zhao. Linvt: Empower your image-level large language model to understand videos. *arXiv preprint arXiv:2412.05185*, 2024. 4
- [11] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. *arXiv preprint arXiv:2410.01615*, 2024. 2, 5, 6
- [12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1, 2, 6
- [13] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. *arXiv preprint arXiv:2405.13382*, 2024. 2, 3, 8
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5, 1
- [15] 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 Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 1, 2, 3, 5, 6, 7, 8
- [16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 2, 6, 7
- [17] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34: 11846–11858, 2021. 6, 7
- [18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3, 4, 7, 8
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3
- [20] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1
- [21] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2, 3, 7
- [22] Xiang Li, Wenhui Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense

- object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 2, 4, 5
- [23] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 3
- [24] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Juntong Pan, Zefeng Li, Van Tu Vu, et al. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024. 1, 3
- [25] Zhiheng Li, Yujie Zhong, Ran Song, Tianjiao Li, Lin Ma, and Wei Zhang. Detal: Open-vocabulary temporal action localization with decoupled networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):7728–7741, 2024. 1
- [26] 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 International Conference on Computer Vision*, pages 2794–2804, 2023. 8
- [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3
- [28] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 7, 2
- [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1
- [30] Weiheng Lu, Jian Li, An Yu, Ming-Ching Chang, Shengpeng Ji, and Min Xia. Llava-mr: Large language-and-vision assistant for video moment retrieval. *arXiv preprint arXiv:2411.14505*, 2024. 8, 2
- [31] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 8
- [32] Boris Meinardus, Anil Batra, Anna Rohrbach, and Marcus Rohrbach. The surprising effectiveness of multimodal large language models for video moment retrieval. *arXiv preprint arXiv:2406.18113*, 2024. 2, 6, 8
- [33] WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023. 2
- [34] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024. 1
- [35] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yuetong Zhuang, and Siliang Tang. Momen-tor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024. 2, 3, 5, 6, 7, 8
- [36] Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. Chatvtg: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2024. 8
- [37] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 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, 2024. 3, 8, 2
- [38] Michael S Ryoo, Honglu Zhou, Shrikant Kendre, Can Qin, Le Xue, Manli Shu, Silvio Savarese, Ran Xu, Caiming Xiong, and Juan Carlos Niebles. xgen-mm-vid (blip-3-video): You only need 32 tokens to represent a video even in vlms. *arXiv preprint arXiv:2410.16267*, 2024. 4
- [39] Dingfeng Shi, Yujie Zhong, Qiong Cao, Jing Zhang, Lin Ma, Jia Li, and Dacheng Tao. React: Temporal action detection with relational queries. In *European conference on computer vision*, pages 105–121. Springer, 2022. 1
- [40] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. *arXiv preprint arXiv:2309.05590*, 2023.
- [41] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18857–18866, 2023. 1
- [42] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 5
- [43] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 6
- [44] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xianguyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 2, 6, 8
- [45] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 7
- [46] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 1, 2, 7, 8
- [47] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024. 1
- [48] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via

- frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022. [8](#)
- [49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. [1](#)
- [50] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. [8](#)
- [51] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. [2](#)
- [52] Yingsen Zeng, Yujie Zhong, Chengjian Feng, and Lin Ma. Unimd: Towards unifying moment retrieval and temporal action detection. In *European Conference on Computer Vision*, pages 286–304. Springer, 2024. [2](#), [6](#), [8](#)
- [53] 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. [1](#), [3](#)
- [54] Pengcheng Zhao, Zhixian He, Fuwei Zhang, Shujin Lin, and Fan Zhou. Ld-detr: Loop decoder detection transformer for video moment retrieval and highlight detection. *arXiv preprint arXiv:2501.10787*, 2025. [6](#)
- [55] Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, pages 20–37. Springer, 2024. [2](#), [6](#)
- [56] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. [1](#), [2](#), [7](#)
- [57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [1](#), [3](#)