iMOVE : Instance-Motion-Aware Video Understanding

Jiaze Li[1,2], Yaya Shi[1]*, Zongyang Ma[3], Haoran Xu[2], Feng cheng[1]
Huihui Xiao[1], Ruiwen Kang[1], Fan Yang[1], Tingting Gao[1], Di Zhang[1]
[1]Kuaishou Technology [2]Zhejiang University
[3]Institute of Automation, Chinese Academy of Sciences

Abstract

Enhancing the fine-grained instance spatiotemporal motion perception capabilities of Video Large Language Models is crucial for improving their temporal and general video understanding. However, current models struggle to perceive detailed and complex instance motions. To address these challenges, we have made improvements from both data and model perspectives. In terms of data, we have meticulously curated **iMOVE-IT**, the first large-scale instance-motion-aware video instruction-tuning dataset. This dataset is enriched with comprehensive instance motion annotations and spatiotemporal mutual-supervision tasks, providing extensive training for the model's instance-motion-awareness. Building on this foundation, we introduce **iMOVE**, an instance-motion-aware video foundation model that utilizes Event-aware Spatiotemporal Efficient Modeling to retain informative instance spatiotemporal motion details while maintaining computational efficiency. It also incorporates Relative Spatiotemporal Position Tokens to ensure awareness of instance spatiotemporal positions. Evaluations indicate that iMOVE excels not only in video temporal understanding and general video understanding but also demonstrates significant advantages in long-term video understanding.

1 Introduction

Recent advancements in Video Large Language Models (Zhang et al., 2023; Maaz et al., 2024b; Li et al., 2023a; Maaz et al., 2024a; Zhang et al., 2024b), i.e., Video-LLMs, have led to rapid development, significantly enhancing the capture of overall video semantics and achieving remarkable performance in general video understanding tasks (Fu et al., 2024; Li et al., 2024). Furthermore, some works (Ren et al., 2024; Huang et al., 2024a,b; Wang et al., 2024a) have begun to explore

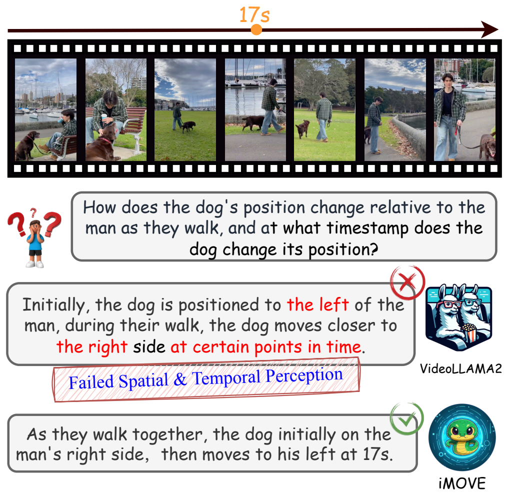

Figure 1: iMOVE excels at perceiving instance-level spatiotemporal motions within videos, outperforming the previous Video-LLM model, VideoLLAMA2.

methods to enable models to comprehend detailed video semantics involving temporal information, showing progress in video temporal understanding tasks (Gao et al., 2017; Caba Heilbron et al., 2015). However, current Video-LLMs struggle to accurately perceive instance spatiotemporal motions within videos, which constrains their performance in these tasks. As demonstrated in Figure 1, VideoLLAMA2 (Zhang et al., 2023) fails to identify the dog's spatiotemporal motions.

The insufficient instance-level motion perception is primarily attributed to two aspects: data and model. On the data side, commonly used training datasets, such as ShareGPT4Video (Chen et al., 2024a), rely on coarse-grained video annotations and lack fine-grained instance spatiotemporal motion annotations. This deficiency makes it challenging for models to accurately perceive instance-level dynamic changes. On the model side, existing Video-LLMs (Li et al., 2023b; Song et al., 2024) do not specifically consider critical spatiotemporal information during the visual token compression process, resulting in an inability to efficiently retain

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 23959–23975
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

all instances' spatiotemporal motions in the video without loss, thereby limiting instance motion perception. Moreover, current works (Ren et al., 2024; Huang et al., 2024b; Wang et al., 2024a) fail to encode both spatial and temporal position information of moving instances in the video simultaneously, leading to weak instance position awareness.

To enable exceptional capture of fine-grained instance motions and achieve superior video understanding, we first construct an **i**nstance-**MO**tion-aware **V**id**E**o **I**nstruction-**T**uning dataset iMOVE-IT, which contains diverse, detailed and rich instance motions. iMOVE-IT defines spatiotemporal mutual-supervision goals, with tasks designed to perform spatial grounding given instance dynamic captions and time information, temporal grounding given instance dynamic captions and spatial information, and instance dynamic captioning given instance time and spatial information. These tasks work in synergy to enhance the model's instance motion awareness, thereby endowing the model with excellent temporal video understanding and general video understanding.

Building on the instance-motion-aware dataset iMOVE-IT, we further introduce iMOVE, a novel **i**nstance-**MO**tion-aware **V**id**E**o foundation model. For visual encoding, iMOVE adopts an Event-aware Spatiotemporal Efficient Modeling strategy, which adaptively segments key events in long videos while preserving the complete spatial appearances of instances and high-frame-rate temporal motions of instances within each event. This ensures the model's detailed understanding of instance motion is not compromised by information loss. For positional encoding, iMOVE introduces Relative Spatiotemporal Position Tokens to simultaneously indicate the spatiotemporal locations of instances during motion, enabling the model to achieve high sensitivity to instance positions. These meticulous designs allow iMOVE to fully understand instance spatiotemporal motions and seamlessly integrate with the iMOVE-IT dataset, unlocking the potential of video understanding.

We conduct comprehensive evaluations on both video temporal understanding and general video understanding benchmarks. The results demonstrate that our approach not only achieves significant improvements in temporal understanding tasks but also excels in general video understanding tasks. Notably, in the zero-shot setting, iMOVE achieves a 10.5% mIOU improvement on the temporal sentence grounding task of Charades-STA (Gao et al., 2017) and a 1.1% SODA_c score improvement on the dense video captioning task of ActivityNet-Captions (Caba Heilbron et al., 2015). Furthermore, in the fine-tuning setting, iMOVE surpasses classical fine-tuned expert models. Meanwhile, iMOVE not only improves general video understanding benchmarks such as MVbench and Video-MME but also significantly enhances long video understanding capabilities. The contributions of this work are listed as follows:

Method	Spatial grounding		Temporal Grounding		Instance Dynamic Captioning	
	Grasp	Generate	Grasp	Generate	Grasp	Generate
Elysium (Wang et al., 2024b)	✗	✓	✓	✗	✓	✓
PiTe (Liu et al., 2024c)	✗	✓	✗	✓	✓	✓
VideoGLaMM (Munasinghe et al., 2024)	✗	✓	✗	✗	✓	✓
INST-IT (Peng et al., 2024)	✓	✗	✓	✗	✓	✓
ViCaS (Athar et al., 2024)	✗	✓	✗	✗	✓	✓
Sa2VA (Yuan et al., 2025)	✗	✓	✗	✗	✓	✓
VideoRefer Suite (Yuan et al., 2024)	✓	✗	✗	✗	✓	✓
iMOVE-IT(Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Instance-Motion-Aware Tasks in Different Studies.

- We collect iMOVE-IT, the first large-scale instance-motion-aware video instruction-tuning dataset with rich instance motions, to improve the model's ability to perceive spatiotemporal motions of instances.
- We propose iMOVE, which employs Event-aware Spatiotemporal Efficient Modeling to encode key instance spatiotemporal motions and introduces Relative Spatiotemporal Position Tokens to represent instances' spatiotemporal positions, thereby acquiring comprehensive instance motion information.
- iMOVE achieves outstanding instance-motion-awareness, leading to significant performance improvements in tasks related to video temporal understanding as well as general and long-term video understanding.

2 Related Work

2.1 Video Large Language Models

With the rapid advancement of large language models (LLMs) and the development of deep learning (Wang and Zhao, 2025b,a, 2024), Video-LLMs have gained significant attention, with early methods like Video-LLAMA (Zhang et al., 2023) and VideoChat (Li et al., 2023a) focusing on capturing overall video semantics through tasks such as video question answering (Maaz et al., 2024b; Zhang et al., 2024b) and video captioning (Chen et al., 2024a; Zhang et al., 2024b), followed by approaches like TimeChat (Ren et al.,

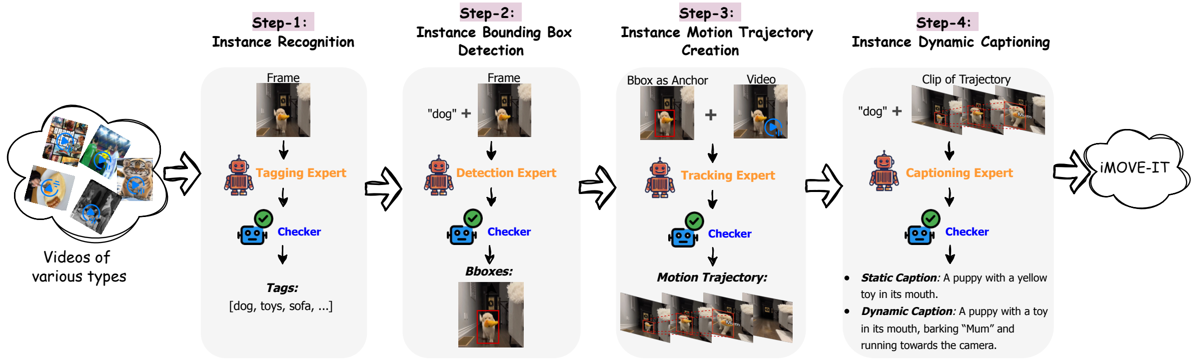

Figure 2: Instance spatiotemporal motion generation pipeline.

2024), VTimeLLM (Huang et al., 2024a), and HawkEye (Wang et al., 2024d) that integrated temporal grounding (Gao et al., 2017), dense video captioning (Caba Heilbron et al., 2015), and highlight detection (Lei et al., 2021) to enhance temporal understanding, while recent multimodal architectures such as Qwen2.5-VL (Bai et al., 2025), Uni-MoE (Li et al., 2025), and Molmo (Deitke et al.) have further expanded Video-LLMs' capabilities through dynamic resolution processing, sparse mixture-of-experts, and novel datasets to improve spatiotemporal reasoning—yet despite these advancements, existing methods still struggle with fine-grained, instance-level comprehension due to limitations in modeling spatiotemporal motion features, which motivates us to propose an instance-level instruction tuning task, iMOVE-IT, and an instance-motion aware model, iMOVE, to address these challenges and advance precise video understanding.

2.2 Instance Perception for Video Understanding

As shown in Table 1, recent methods have enhanced fine-grained spatiotemporal awareness through instance-level supervision tasks, categorized into spatial grounding, temporal grounding, and instance dynamic captioning, and further divided into input-side (grasp) and output-side (generate) based on model placement. Notable contributions include ElysiumTrack-1M (Wang et al., 2024b) for single object tracking, referring tracking, and video referring expression generation; INST-IT (Peng et al., 2024) with its instance-specific instruction tuning dataset focusing on states, transitions, and QA pairs; and VideoRefer-700K (Yuan et al., 2024) providing region-level annotations with detailed captions and multi-round QA pairs for object-level video understanding. Compared to these existing methods involving tasks from either grasping or generating perspectives with limited coverage, our iMOVE-IT dataset provides comprehensive tasks for instance-level spatiotemporal perception.

3 Method

In this section, we outline the construction of iMOVE as follows: collect the dataset iMOVE-IT with an automated pipline in Sec 3.1 and build the model iMOVE in Sec 3.2.

3.1 iMOVE-IT

3.1.1 Instance Spatiotemporal Motion Generation

As shown in Figure 2, we propose an automatical pipeline for generating instance motion trajectories and corresponding dynamic captions. The specific steps are as follows:

Step-1: Instance Recognition The initial frame of the given video is processed by the tagging expert RAM++ (Huang et al., 2023) to identify all object categories present. Human checkers review these object categories and filter out some static categories, such as cloud, sky, road, and room.

Step-2: Instance Bounding Box Detection.

For each recognized category, we employ the open-vocabulary detection expert Grounding DINO (Liu et al., 2024b) to find all relevant instances and their associated bounding boxes. Subsequently, we filter out unreasonable bounding boxes, specifically those that are excessively large or those that do not match the corresponding category names based on CLIP (Radford et al., 2021) similarity scores. The filtering thresholds are optimized based on feedback from human checkers.

Step-3: Instance Motion Trajectory Creation. Each detected instance bounding box is treated as a tracking target and fed into the visual tracking

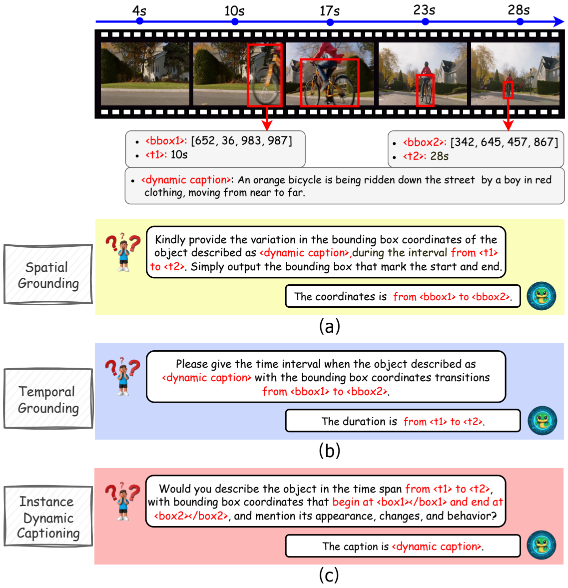

Figure 3: Examples of iMOVE-IT.

expert SAMURAI (Yang et al., 2024), which outputs the instance motion trajectory. We filter out unreasonable trajectories, i.e., those with bounding boxes that differ significantly in area from the initial bounding boxes, and those with low CLIP similarity scores between the bounding boxes and their corresponding category names. The filtering thresholds are determined based on feedback from human checkers who prioritize accuracy.

Step-4: Instance Dynamic Captioning. To enhance the accuracy of data generation, we query the captioning expert InternVL2-40B(Chen et al., 2024c) twice, separately generating static and dynamic captions. Specifically, by combining the category name, bounding box, and motion trajectory of each instance, we first prompt InternVL2-40B to generate accurate static captions that describe the instance's appearance. Following this, based on the static captions, InternVL2-40B generates dynamic captions that describe the spatiotemporal motion of the instance. Additionally, we employ Qwen2-VL-7B-Instruct(Wang et al., 2024c) as a checker to filter out erroneous dynamic captions and those that cannot uniquely refer to the instance, thereby improving the quality of the generated data.

3.1.2 iMOVE-IT Construction

As shown in Figure 3, by utilizing an automatical pipeline for data generation, we constructed iMOVE-IT, which encompasses three tasks: Spatial Grounding, Temporal Grounding, and Instance Dynamic Captioning.

Spatial Grounding. This task involves predicting the bounding box coordinates <bbox1, bbox2> for the relevant instances in the start and end frames, based on the given dynamic caption <dynamic caption> and the time interval <t1, t2>. This process enhances the spatial grounding capability, enabling precise localization of target instances across consecutive video frames.

Temporal Grounding. The task involves predicting the time interval <t1, t2> corresponding to the motion trajectory of an instance based on the provided dynamic caption < dynamic caption> and bounding box coordinates <bbox1, bbox2> at the start and end of the trajectory. This task enhances the model's temporal awareness of video segments, enabling it to determine time intervals corresponding to specific dynamic captions and changes in spatial positions.

Instance Dynamic Captioning. The task aims to generate a dynamic caption based on the given time interval <t1, t2> and the bounding box coordinates <bbox1, bbox2>, which represent the start and end of the instance's motion trajectory within this time interval. This task strengthens the model's ability to match video and text, enabling it to generate dynamic caption that accurately reflect the specific instance's spatiotemporal motion within the given time interval.

The proposed iMOVE-IT dataset consists of three meticulously designed instance-level spatiotemporal mutual-supervision tasks, which provide abundant spatiotemporal supervisory signals to enhance the model's fine-grained instance spatiotemporal motion perception capabilities from the data perspective.

3.1.3 Statistics of the iMOVE-IT Dataset

iMOVE-IT comprises 114k video-instruction pairs from 68k unique videos, with an average duration of 109 seconds. To ensure the integrity of zero-shot setting, the dataset excludes videos from ActivityNet (Caba Heilbron et al., 2015) and Charades-STA (Gao et al., 2017), while incorporating data from 11 distinct datasets. Detailed statistics are provided in Appendix A.1.

3.2 iMOVE

As shown in Figure 4, iMOVE adapts short Video-LLMs to perceive fine-grained spatiotemporal instance motions, thereby enhancing its temporal and general video understanding capabilities, while improving long-term video understanding.

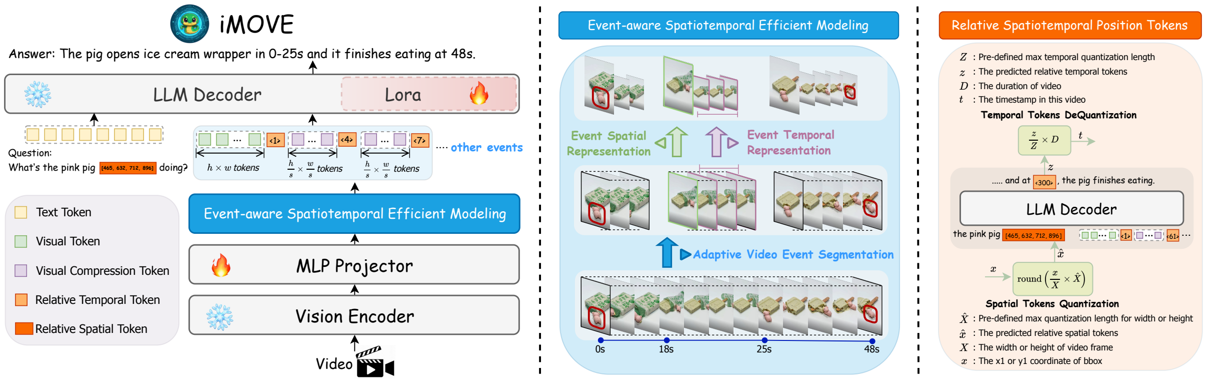

Figure 4: Architecture of iMOVE. iMOVE employs Event-aware Spatiotemporal Efficient Modeling to encode long videos while preserving key instance motions. Additionally, iMOVE introduces Relative Spatiotemporal Positional Token, enhancing sensitivity to spatiotemporal locations of instance motions within the video.

For a long video, we sample T frames and encode each into N visual tokens using a visual encoder. These tokens are projected into the language semantic space, creating a feature map $\mathbf{F} \in \mathbb{R}^{\hat{h}\times\hat{w}\times d}$ for each frame, where $\hat{h}$ and $\hat{w}$ represent the dimensions of the feature map, and d is the token dimension size. The complete video is represented as $\mathbf{V} = (\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_T) \in \mathbb{R}^{T\times\hat{h}\times\hat{w}\times d}$.

We further propose **Event-aware Spatiotemporal Efficient Modeling**, which reduces visual tokens in dense video encoding while preserving essential spatiotemporal motion, balancing model efficiency and video representation integrity. We also propose **Relative Spatiotemporal Position Tokens** to enhance the model's understanding of instances' spatiotemporal positions. These innovations will be presented in the subsequent sections.

3.2.1 Event-aware Spatiotemporal Efficient Modeling

For the raw dense visual features $\mathbf{V}$, we propose the Event-aware Spatiotemporal Efficient Modeling method, which maintains computational efficiency while preserving the complete spatial appearances of instances and the high-frame-rate temporal motions of instances. Specifically, iMOVE first adaptively segments the video into events. The frame with the most information in each event is identified as the one retaining the most spatial information. Subsequently, other frames with less information in the event undergo greater spatial compression to preserve the high-frame-rate detailed temporal and motion information of the instance. Notably, this computationally efficient process does not introduce new parameters, thereby reducing the overall model training complexity.

Adaptive Video Event Segmentation. We propose a novel adaptive video event segmentation method that captures event transitions in videos by identifying significant changes in inter-frame similarity. First, we calculate the cosine similarity between adjacent frame features:

$$S_{i,i+1} = \frac{\mathbf{F}_i \cdot \mathbf{F}_{i+1}}{\|\mathbf{F}_i\|\|\mathbf{F}_{i+1}\|} \tag{1}$$

where $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$ are the feature vectors of adjacent frames. Then, we compute the rate of change in similarity:

$$d_i = |S_{i,i+1} - S_{i-1,i}| \tag{2}$$

Next, we sort d_i in descending order and select the largest $K-1$ extreme points as event boundaries, thereby segmenting the video into K events.

Event Spatial Representation. According to information entropy theory, adjacent frames with larger similarity differences carry more information, while those with smaller differences contain redundancy. Based on this, we consider the first frame of each event, which occurs at the moment with the highest rate of information change, to be the frame with the most instance appearance information of the event and designate it as the event's spatial representation.

Event Temporal Representation. Non-first frames in each event provide less spatial information than the first frame but contain rich temporal information. Therefore, we use average pooling to merge dense visual tokens spatially. Assuming the first frame feature in each event is $\hat{\mathbf{F}} \in \mathbb{R}^{h\times w\times d}$, subsequent non-first frame features are compressed based on this dimension. Using s as stride, each non-first frame feature is pooled into $\hat{\mathbf{F}} \in \mathbb{R}^{\frac{h}{s}\times\frac{w}{s}\times d}$, preserving high-frame-rate temporal motions. This strategy allows the model to

Model	LLM	Charades-STA				ActivityNet-Grounding				ActivityNet-Captions	
	Scale	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU	SODA_c	METEOR
Zero-Shot											
Video-ChatGPT (Maaz et al., 2024b)	7B	27.2	6.2	1.9	19.7	19.5	10.6	4.8	14.2	1.9	2.1
VideoChat (Li et al., 2023a)	7B	32.8	8.6	0.0	25.9	23.5	12.6	6.0	17.4	0.9	0.9
Momentor (Qian et al., 2024)	7B	42.6	26.6	11.6	28.5	**42.9**	23.0	**12.4**	29.3	2.3	4.7
TimeChat (Ren et al., 2024)	7B	-	32.2	13.4	-	-	-	-	-	-	-
VTG-LLM (Guo et al., 2024a)	7B	-	33.8	15.7	-	-	-	-	-	-	-
HawkEye ★ (Wang et al., 2024d)	7B	50.6	31.4	14.5	33.7	49.1	29.3	10.7	32.7	-	-
PiTe★ (Liu et al., 2024c)	7B	-	-	-	-	30.4	17.8	7.8	22.0	5.1	5.8
Grounded-VideoLLM ★ (Wang et al., 2024a)	4B	54.2	36.4	19.7	36.8	46.2	30.3	19.0	36.1	6.0	6.8
VTimeLLM ★ (Huang et al., 2024a)	7B	51.0	27.5	11.4	31.2	44.0	27.8	14.3	30.4	5.8	6.8
TIMESUITE (Zeng et al., 2024b)	7B	69.9	48.7	24.0	-	-	-	-	-	-	-
TRACE (Guo et al., 2024b)	7B	-	40.3	19.4	-	-	-	-	-	-	-
InternVL2-4B(Chen et al., 2024c)	4B	14.7	7.9	3.3	11.9	17.1	9.5	4.6	12.7	0.85	2.81
iMOVE	4B	**71.7**	**51.3**	**26.1**	**47.3**	42.4	**23.1**	12.1	**29.7**	**3.4**	**6.8**
Fine-Tuning											
Vid2Seq ♠(Yang et al., 2023)	-	-	-	-	-	-	-	-	-	5.8	-
QD-DETR♠(Moon et al., 2023)	-	-	57.3	32.6	-	-	-	-	-	-	-
UnLoc-L ♠ (Yan et al., 2023)	-	-	60.8	38.4	-	-	48.3	30.2	-	-	-
VTG-LLM (Guo et al., 2024a)	7B	-	57.2	33.4	-	-	-	-	-	-	-
HawkEye (Wang et al., 2024d)	7B	72.5	58.3	28.8	49.3	55.9	34.7	17.9	39.1	-	-
TIMESUITE (Zeng et al., 2024b)	7B	79.4	67.1	43.0	-	-	-	-	-	-	-
TRACE (Guo et al., 2024b)	7B	-	61.7	41.4	-	-	37.7	24.0	39.0	6.0	6.4
iMOVE+FT	4B	**79.8**	**68.5**	**45.3**	**57.9**	**67.2**	**50.7**	**32.4**	**49.3**	**6.0**	**8.0**

Table 2: Zero-Shot and Fine-Tuning results on Temporal Sentence Grounding and Dense Video Captioning tasks. **Bold** fonts highlight the best performance. Underline highlights the second best performance. Fine-tuned expert models are marked with ♠, while non-strict zero-shot methods on ActivityNet-Captions dataset are marked with ★.

process more temporal frames without increasing the total visual token length, effectively modeling high-frame-rate event temporal representation.

3.2.2 Relative Spatiotemporal Position Tokens

Taking temporal tokens as examples, existing methods can be categorized into absolute and relative time representations. The former, such as TimeChat (Ren et al., 2024) and TimeMarker (Chen et al., 2024b), use absolute time tokens like "2s" or "Second8.0". However, the broad nature of absolute time ranges and the impossibility of exhaustive enumeration make it challenging to generalize across videos of varying lengths. The latter, such as LITA (Huang et al., 2024b) and Grounded-VideoLLM (Wang et al., 2024a), introduce special time tokens into the tokenizer of LLMs. However, this requires modifying word embedding parameters, potentially disrupting their compatibility with the LLM's parameters and risking performance degradation, especially in small-scale fine-tuning scenarios. Therefore, in this paper, we utilize relative spatiotemporal tokens to encode bounding box coordinates and timestamp positions without adding them into the tokenizer. Unlike previous methods using relative time representations (Wang et al., 2024a; Huang et al., 2024b), we append a relative temporal token indicating the corresponding time to the visual tokens of each frame. These tokens are interleaved with pruned visual tokens, enabling the model to more accurately perceive the temporal information of the video. Specifically, we use quantization and dequantization on input and output sides to map actual values to relative tokens.

For temporal token, given a video of duration D seconds, we establish a bidirectional mapping between the timestamp t and a discrete value $z \in [0, Z]$, where Z is an empirical value. The quantization process is:

$$z = \text{round}\left(\frac{t}{D} \times Z\right) \tag{3}$$

and the dequantization process is:

$$t = \frac{z}{Z} \times D \tag{4}$$

For spatial token, given a frame with width W and height H, we establish a bidirectional mapping between the x coordinate and a discrete value $\hat{x} \in [0, \hat{W}]$, as well as between the y coordinate and a discrete value $\hat{y} \in [0, \hat{H}]$, in the same way as temporal token.

4 Experiments

The detailed experimental setup and hyperparameters can be found in Appendix B.1. Baseline and comparison details are provided in Appendix B.2. We have performed rigorous data filtering to ensure the zero-shot setting on Charades-STA and ActivityNet-Captions datasets.. Detailed

Model	LLM	MVBench			Video-MME(w/o subs)		LongVideoBench
	Scale	AS	AP	Avg	Long	Overall	val
Video-LLaMA (Zhang et al., 2023)	7B	27.5	25.5	34.1	-	-	-
Video-ChatGPT (Maaz et al., 2024b)	7B	23.5	26.0	32.7	-	-	-
VideoChat2 (Li et al., 2024)	7B	66.0	47.5	51.1	33.2	39.5	-
ST-LLM (Liu et al., 2024a)	7B	66.0	53.5	54.9	31.3	37.9	-
VideoGPT+ (Maaz et al., 2024a)	7B	69.0	60.0	58.7	-	-	-
MovieChat (Song et al., 2024)	7B	-	-	55.1	33.4	38.2	-
P-LLaVA-13B (Xu et al., 2024)	13B	66.0	53.0	50.1	-	-	45.6
LLaVA-Next-Video-34B (Zhang et al., 2024a)	34B	-	-	-	-	-	50.5
TIMESUITE (Zeng et al., 2024b)	7B	-	-	59.9	41.9	46.3	-
TRACE (Guo et al., 2024b)	7B	-	-	48.1	-	43.8	-
InternVL2-4B(Chen et al., 2024c)	4B	**76.0**	62.5	63.7	39.2	48.7	50.7
iMOVE	4B	71.5	**62.5**	**63.9**	**43.2**	**53.6**	**54.7**

Table 3: Zero-Shot results on General Video Understanding and Long-term Video Understanding tasks. **Bold** fonts highlight the best performance. Underline highlights the second best performance.

training data composition and data filtering procedures are described in Appendix B.3. Qualitative analysys are available in Appendix C.

4.1 Main Results

To comprehensively assess the video understanding capabilities, we conducted quantitative evaluations across three task categories: Video Temporal Understanding(including Temporal Video Grounding and Dense Video Captioning), General Video Understanding and Long-term Video Understanding. Details of the benchmarks and evaluation metrics refer to Appendix B.4. Notably, PiTe(Liu et al., 2024c), Grounded-VideoLLM(Wang et al., 2024a), HawkEye(Wang et al., 2024d), and VTimeLLM(Huang et al., 2024a) do not operate under a strict zero-shot setting on the ActivityNet-Captions dataset. Detailed explanations can be found in Appendix B.5.

Temporal Video Grounding. This task aims to identify the start and end timestamps of events described by a given query sentence. As shown in Table 2, under the zero-shot setting, iMOVE achieves mIoU accuracies of 47.3% and 29.7% on Charades-STA and ActivityNet-Grounding, surpassing previous SOTAs, i.e., Grounded-VideoLLM and Momentor, by margins of 10.5% and 0.4%. With fine-tuning, iMOVE further attains mIoU accuracies of 57.9% and 49.3% on these benchmarks, significantly outperforming existing SOTA methods by 10.6% and 10.2%. This highlights iMOVE's superior fine-grained temporal localization capability.

Dense Video Captioning. This task requires detecting all events in videos while providing corresponding duration time intervals and descriptions. As observed in Table 2, iMOVE achieves SODA_c (Fujita et al., 2020) and METEOR (Banerjee and Lavie, 2005) scores of 3.4 and 6.8 on ActivityNet-Captions under the zero-shot setting, exceeding the prior SOTA method Momentor (Qian et al., 2024) with scores of 2.3 and 4.7. After fine-tuning, iMOVE also outperforms the specialized model TRACE (Guo et al., 2024b). The improvements can be attributed to iMOVE's meticulously designed key-event retention strategy, which enables accurate capture of complete event narratives.

General Video Understanding. Video-MME (Fu et al., 2024) and MVBench (Li et al., 2024) are used to evaluate general video understanding capabilities. According to the results in Table 3, despite utilizing fewer visual tokens, iMOVE obtains accuracies of 53.6 and 63.9 on Video-MME and MVBench, improving by 4.9% and 0.2% compared to InternVL2-4B. While prior temporal-focused MLLMs excel at time-related tasks, they exhibit compromised performance on general video understanding tasks. In contrast, iMOVE performs best on both tasks, demonstrating that enhanced fine-grained instance motion modeling synergistically benefits dual capabilities.

Long-term Video Understanding. The Long Video subset of Video-MME and LongVideoBench (Wu et al., 2024) serve as benchmarks for long video understanding evaluation. The comparisons are listed in Table 3, iMOVE acquires accuracies of 54.7 on LongVideoBench and 43.2 on Video-MME Long, both of which represent a 4% improvement over the similarly parameter-sized Interval2-4B. Additionally, iMOVE significantly outperforms Video-LLMs with larger parameter scales, e.g., LLaVA-Next-Video-34B (Zhang et al., 2024a) and P-LLaVA-13B (Xu et al., 2024). This verifies that strengthening instance motion perceiving improves long video understanding.

DVC+TAL	TSG	Geneal	iMOVE-IT	C-STA mIoU	ANet-G mIoU	ANet-Cap SODA_c	LVBench val	V-MME Overall	MVB Avg
✓				25.4	20.3	Fail	46.6	47.0	57.0
✓	✓			44.8	27.8	1.8	50.7	50.3	57.2
✓	✓	✓		46.1	27.5	2.8	53.6	53.8	63.4
✓	✓	✓	✓	47.3	29.7	3.4	54.7	53.6	63.9

Table 4: Ablations on training data composition. DVC, TAL, and TSG stands datasets for Dense Video Captioning, Temporal Action Localization and Temporal Sentence Grounding tasks. General includes datasets from three tasks: VideoQA, Classification, and Video Captioning. Fail is an inability to follow instructions.

Row	Rand	Uniform	Event	K	s	# of tokens	C-STA mIoU	V-MME Overall
1	✓			24	2	2688	43.2	50.0
2		✓		24	2	2688	43.6	49.9
3			✓	24	2	2688	44.8	50.3
4			✓	0	2	1536	43.8	49.2
5			✓	12	2	2112	45.0	49.8
6			✓	48	2	3840	45.1	51.3
7			✓	96	2	6144	46.2	51.1
8			✓	24	1	6144	46.2	51.1
9			✓	24	4	1824	44.9	49.9
10			✓	24	8	1608	44.2	49.1

Table 5: Effect of Event-aware Spatiotemporal Efficient Modeling. Rand denotes randomly selecting the first frame of the Event, Uniform denotes uniformly selecting frames, and Event denotes our method.

4.2 Ablation Study

Benefits of iMOVE-IT. As shown in Table 4, the model performance steadily improved with the incremental inclusion of datasets from different tasks. The row 4 vs. row 3 comparison reveals that incorporating iMOVE-IT significantly enhances instance motion awareness, leading to improved temporal understanding, long-term video understanding, and general video understanding capabilities.

Effectiveness of Event-aware Spatiotemporal Efficient Modeling. The first three rows of Table 5 demonstrate that our proposed Event-aware Spatiotemporal Efficient Modeling outperforms both random and uniform frame selection strategies. Furthermore, as observed from rows 3 to 5, increasing the number of segmented events K leads to continuous improvements on Charades-STA and Video-MME datasets. Rows 8 to 10 indicate that as the non-first frame pooling rate s increases, there is a loss of spatial information, which results in a decline in the model's temporal understanding and generalization capabilities. Conversely, when the pooling rate is too small ($s = 1$), the excessive number of tokens adversely affects the model's training and inference efficiency. Balancing information loss and efficiency, we selected a pooling rate of ($K = 24$) and ($s = 2$) as the optimal trade-off. The composition of the data for this ablation study can be found in Appendix B.6.

Row	RT	AT	Add-VT	Add-Tokenizer	RST	C-STA mIoU	ANet-G mIoU
1	✓					44.9	26.3
2	✓		✓			46.6	29.2
3	✓		✓		✓	**47.3**	**29.7**
4		✓	✓		✓	45.7	28.4
5	✓		✓	✓	✓	44.4	27.6

Table 6: Effect of Relative Spatiotemporal Position Tokens: RT and AT denote relative and absolute temporal representations. Add-VT appends the temporal token after each frame's visual tokens, while Add-Tokenizer incorporates temporal position tokens into the tokenizer. RST represents Relative Spatial Position Token.

Architecture	C-STA	ANet-G	ANet-Cap		LVBench
	mIoU	mIoU	SODA_c	METEOR	val
Phi3.5-V	45.5	29.2	3.3	7.2	53.4
InternVL2-4B	47.3	29.7	3.4	6.8	54.7

Table 7: Generalization Study. Phi3.5-V denotes Phi3.5-Vision-Instruct-3.8B.

About of Relative Spatiotemporal Position Tokens. Table 6 shows that row 2 and row1 indicate that adding corresponding temporal token after each frame's visual token is beneficial for achieving better model performance. The performance improvement of row 3 compared to row 4 demonstrates that relative temporal representation outperforms absolute temporal representation in our experiments. The performance decline of row 5 relative to row 3 suggests that introducing additional temporal markers in the LLM's tokenizer results in suboptimal outcomes. Meanwhile, the performance improvement of row 3 compared to row 2 proves the effectiveness of the Relative Spatial Position Token. In summary, these findings demonstrate the effectiveness of the Relative Spatiotemporal Position Tokens.

Generalization Study. To validate the generalization capability of the proposed method, we employ an image MLLM, Phi3.5-Vision-Instruct-3.8B (Microsoft., 2024), as our base MLLM. As shown in Table 7, using an image MLLM as the base achieves comparable performance to a short VideoLLM. Notably, it improves the METEOR score by 0.4 points on ActivityNet-Captions. This indicates that our model architecture and dataset not only enable short Video-LLMs to adapt to perceiving fine-grained spatiotemporal instance motions but are also applicable to image MLLMs.

5 Conclusion

In this paper, we enhanced the fine-grained instance spatiotemporal motion perception in Video-LLMs by making improvements from both data and model perspectives, thereby boosting their capabilities in temporal and general video understanding. Data-wise, we propose **iMOVE-IT**, a large instance-motion-aware video dataset with spatiotemporal mutual-supervision tasks to enhance instance spatiotemporal motion learning. Model-wise, we develop **iMOVE** featuring Event-aware Spatiotemporal Efficient Modeling for efficiently preserving motion details, and Relative Spatiotemporal Position Tokens for accurate spatial-temporal positioning.

Evaluations indicate that iMOVE excels in video temporal understanding and general video understanding, with significant advantages in long-term video understanding, offering valuable insights for video understanding research.

6 Limitations

The selection of the number of events K in Event-aware Spatiotemporal Efficient Modeling is fixed. Although we believe that choosing 24 can cover the vast majority of videos and experiments have demonstrated its effectiveness, adapting the number of events for different videos is a promising way to enhance our method. We will explore this aspect in our future work. Secondly, the iMOVE-IT dataset proposed in this paper has achieved excellent results by enhancing the model's fine-grained instance spatiotemporal motion perception capabilities, thereby improving the model's temporal understanding and general video understanding. However, our primary focus is on the fine-grained spatiotemporal understanding task of bounding box coordinates. Exploring more fine-grained spatiotemporal tasks is a promising direction for future research.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Ali Athar, Xueqing Deng, and Liang-Chieh Chen. 2024. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. *arXiv preprint arXiv: 2412.09754*.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024b. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv: 2411.18211*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. `https://internvl.github.io/blog/2024-07-02-InternVL-2.0/`.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

M Deitke, C Clark, S Lee, R Tripathi, Y Yang, JS Park, M Salehi, N Muennighoff, K Lo, L Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. *URL https://arxiv. org/abs/2409.17146*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 517–531. Springer.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*, pages 5267–5275.

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

Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. 2024b. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv: 2410.05643*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024a. Vtimellm: Empower llm to grasp video moments. In *CVPR*.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. 2024b. Lita: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. 2023. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310.

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.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, pages 11846–11858.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024a. St-llm: Large language models are effective temporal learners. In *ECCV*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55.

Yang Liu, Pengxiang Ding, Siteng Huang, Min Zhang, Han Zhao, and Donglin Wang. 2024c. Pite: Pixel-temporal alignment for large video-language model. In *European Conference on Computer Vision*, pages 160–176. Springer.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024a. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024b. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*.

Microsoft. 2024. Phi-3.5 vision instruct. https://huggingface.co/microsoft/Phi-3.5-vision-instruct.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. 2023. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23023–23033.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2024. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv: 2411.04923*.

Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2024. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning. *arXiv preprint arXiv: 2412.03565*.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *ICML*.

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 *CVPR*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

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

Hang Wang, Yanjie Wang, Yongjie Ye, Yuxiang Nie, and Can Huang. 2024b. Elysium: Exploring object-level perception in videos via mllm. *ECCV*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv: 2409.12191*.

Yifeng Wang and Yi Zhao. 2024. Chinese inertial gan for writing signal generation and recognition.

Yifeng Wang and Yi Zhao. 2025a. General pre-trained inertial signal feature extraction based on temporal memory fusion. *Information Fusion*, page 103274.

Yifeng Wang and Yi Zhao. 2025b. Heros-gan: Honed-energy regularized and optimal supervised gan for enhancing accuracy and range of low-cost accelerometers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21348–21356.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024d. Hawkeye: Training video-text llms for grounding text in videos. *arXiv preprint arXiv:2403.10228*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. 2025. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. 2023. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, pages 10714–10726.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. 2024. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint arXiv: 2411.11922*.

Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. 2025. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv: 2501.04001*.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. 2024. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *arXiv preprint arXiv: 2501.00599*.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. 2024a. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv: 2410.19702*.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. 2024b. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*.

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*.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024a. Llava next video. `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *arXiv preprint arXiv: 2410.02713*.

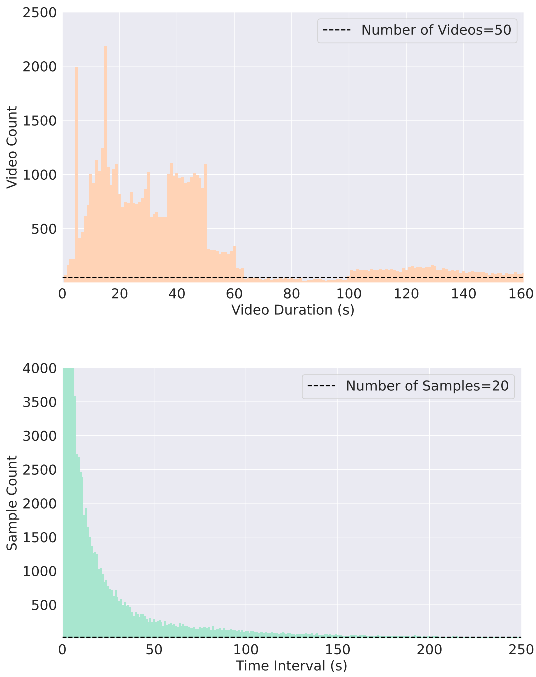

Figure 5: (a) Video duration visualization for iMOVE-IT. (b) Visualization of time intervals in iMOVE-IT.

A More Details of iMOVE-IT

A.1 Data Statics of iMOVE-IT

Diversity of Video Types The iMOVE-IT dataset contains 68387 unique videos from 11 different datasets, including InternVid-10M, CLEVRER, DiDeMo, NExT-QA, HACS, YT-Temporal-1B, COIN, TACoS, YouCook2, HiREST and ViTT, and does not include videos from the ActivityNet and Charades-STA datasets. The average duration of the videos is 109 seconds. The distribution of video durations in iMOVE-IT is shown in Figure 5(a). As can be seen from the figure, the distribution of video durations is quite broad, with the majority of videos being less than 60 seconds long. Videos longer than 60 seconds are primarily concentrated in the range of 100 to 160 seconds.

Diversity of Object Types The iMOVE-IT dataset encompasses 114,705 objects, derived from 23,278 object categories, with an average of 5 objects per category. Specifically, the Spatial grounding task consists of 51,992 objects, the Temporal Grounding task includes 21,328 objects, and the Instance Dynamic Captioning task comprises 41,385 objects.

Diversity of time intervals Figure 5(b) depicts the distribution of time intervals in iMOVE-IT, with the majority of time intervals concentrated within 50 seconds. This is likely due to the fact that the video durations are primarily between 1 and 60 seconds. Additionally, the diversity in the distribution of time intervals contributes to enhancing the model's robustness.

From the above analysis, we can see that the iMOVE-IT dataset has significant advantages in terms of sample size, video types, object types, time intervals, and task instruction diversity. These characteristics make this dataset highly valuable for research and applications in related fields.

A.2 Task Prompts of iMOVE-IT

Prompts of iMOVE-IT

Spatial Grounding

1. Please give the bounding box coordinates variation of the object depicted as `<dynamic caption>` during the time interval from `<t1>` to `<t2>`. Output only the bounding box coordinates for the start and end times.
2. I would like to know the bounding box coordinates variation of the object described as `<dynamic caption>` during the period from `<t1>` to `<t2>`. Just produce the bounding box coordinates for when it starts and ends.
3. Can you give me the bounding box coordinates change for the object depicted as `<dynamic caption>` from `<t1>` to `<t2>`? Only the bounding box coordinates for the beginning and conclusion should be generated.
4. Kindly provide the variation in the bounding box coordinates of the object described as `<dynamic caption>` from `<t1>` to `<t2>`. Simply output the bounding box coordinates that mark the start and end.
5. What is the change in the bounding box coordinates of the object depicted as `<dynamic caption>` during the interval from `<t1>` to `<t2>`? Provide solely the bounding box coordinates that denote the start and end points.
6. Let me know the variation in the bounding box coordinates of the object described as `<dynamic caption>` during the time span from `<t1>` to `<t2>`. Ensure the output includes only the bounding box coordinates for the start and end.
7. I need the bounding box coordinates change of the object depicted as `<dynamic caption>` during the period from `<t1>` to `<t2>`. Limit the output to the bounding box coordinates of the start and end times.
8. Kindly tell me the change in the bounding box coordinates for the object depicted as `<dynamic caption>` between `<t1>` and `<t2>`. You need to output solely the bounding box coordinates for the start and end times.

Temporal Grounding

1. Please give the time interval when the object described as `<dynamic caption>` transitions from the bounding box coordinates `<bbox1>` to `<bbox2>`.
2. Could you provide the time interval for the transition of the object, described as `<dynamic caption>`, from the bounding box coordinates `<bbox1>` to `<bbox2>`?
3. What is the time interval when the object, depicted as `<dynamic caption>`, shifts from the bounding box `<bbox1>` to `<bbox2>`?
4. Kindly indicate the time span for the object, described as `<dynamic caption>`, transitioning from `<bbox1>` to `<bbox2>`.
5. Please let me know the duration of time when the object, with a description of `<dynamic caption>`, moves from the coordinates `<bbox1>` to `<bbox2>`.
6. Identify the time interval for the object, described as `<dynamic caption>`, as it transitions from `<bbox1>` to `<bbox2>`.
7. Could you specify the time period for the object, with a depiction of `<dynamic caption>`, moving from `<bbox1>` to `<bbox2>`?
8. Please tell me the duration of time for the object, with a portrayal of `<dynamic caption>`, as it transitions from `<bbox1>` to `<bbox2>`.

Prompts of iMOVE-IT (continued)

Instance Dynamic Captioning

1. Please describe the object within the time interval from <t1> to <t2>, with bounding box coordinates starting at <bbox1> and ending at <bbox2>. Include the object's appearance, changes, and behavior.
2. Could you describe the object within the time interval from <t1> to <t2>, with bounding box coordinates beginning at <bbox1> and ending at <bbox2>, and include its appearance, changes, and behavior?
3. Describe the object from <t1> to <t2>, with bounding box coordinates that begin at <bbox1> and end at <bbox2>, and include information on its appearance, alterations, and behavior.
4. Could you give a description of the object that is positioned between <t1> and <t2>, with bounding box coordinates commencing at <bbox1> and finishing at <bbox2>, and include its appearance, changes, and behavior?
5. Would you describe the object in the time span from <t1> to <t2>, with bounding box coordinates that begin at <bbox1> and end at <bbox2>, and mention its appearance, changes, and behavior?
6. Please provide a description of the object between <t1> and <t2>, with bounding box coordinates starting at <bbox1> and ending at <bbox2>, and include its appearance, changes, and behavior.
7. Describe the object from <t1> to <t2>, with bounding box coordinates initiating at <bbox1> and concluding at <bbox2>, and incorporate its appearance, variations, and behavior.
8. Could you portray the object from <t1> to <t2>, with bounding box coordinates that commence at <bbox1> and finish at <bbox2>, and include its appearance, changes, and behavior?

B EXPERIMENTS

B.1 Detailed Dxperiment Setup

Learning Rate	1e-4
LR Scheduler	Cosine
Global Batch Size	128
Training Steps	10K
Warmup Ratio	0.03
Trainable Modules	MLP Projector & LoRA
Frame Resolution	448×448
Num Frames	96
Train Epochs	1
Model Max Length	10000
ZeRO Optimization	Zero-2
Computation	16 A800

Table 8: The hyper-parameters for iMOVE.

We utilize InternViT-300M-448px (Chen et al., 2024d) as the video encoder and Phi-3-mini-128k-instruct (Abdin et al., 2024) as the LLM decoder. The parameters of equivalent components from InternVL2-4B, including linear projection layers, are used to initialize these models. The maximum value for temporal token quantization, denoted as Z, is set to 300. For spatial tokens, the maximum quantization values for height $\hat{H}$ and width $\hat{W}$ are both set to 1000. We fine-tune the LLM using LoRA (Hu et al., 2022), while keeping the visual encoder frozen and making the MLP trainable. The LoRA parameters are set to $r = 128$ and $\alpha = 256$. We employ the AdamW (Loshchilov and Hutter, 2019) optimizer with a warm-up rate of 0.03. Initially, the video is divided into 96 segments, and one frame is randomly selected from each segment, resulting in 96 video frames. For each video, we select $K = 24$ events, setting the first frame feature's h and w of each event to 8, thus encoding with 64 tokens. The stride s is set to 2, meaning that non-first frames in each event are encoded with 16 tokens, resulting in a total of 2688 tokens per video, which is significantly fewer than the 4096 tokens per video encoding method in InternVL2-4B. Additional hyper-parameters can be found in Table 8. All experiments are conducted on 16 A800 GPUs with a batch size of 128. The training set we used contains 1,276,977 data samples, and the experiment took a total of 41 hours to complete. Furthermore, in the construction of iMOVE-IT, we utilized the ViT-B/32 version of CLIP.

B.2 Baseline and Comparison

For the baseline models, the comparison primarily involves two categories: the first is the zero-shot video large language models, and the second is the expert models obtained through fine-tuning. For the first type of baseline method, we choose InternVL2-4B(Chen et al., 2024c), Video-ChatGPT (Maaz et al., 2024b), VideoChat (Li et al., 2023a), Momentor (Qian et al., 2024), TimeChat (Ren et al., 2024), VTG-LLM (Guo et al., 2024a), HawkEye (Wang et al., 2024d), PiTe (Liu et al., 2024c), Grounded-VideoLLM (Wang et al., 2024a), VTimeLLM (Huang et al., 2024a), TimeSuite (Zeng et al., 2024a), and TRACE (Guo et al., 2024b). We selected Vid2Seq(Yang et al., 2023) as the classic supervised expert model on the Charades-STA dataset. For the ActivityNet-Captions, we chose QD-DETR(Moon et al., 2023) and UnLoc-L(Yan et al., 2023) as the supervised expert models. Additionally, we report the results of some zero-shot models after fine-tuning.

B.3 Detailed Data Filtering and Dataset Composition

In terms of data quality control, we implemented stringent measures. We excluded datasets such as STAR, ANet-RTL, VCG-Plus112K, Videochatgpt-100K, Videochat2-Conv, and TextVR to ensure a strict zero-shot setting on the Charades-STA and ActivityNet-Captions datasets. Additionally, the LLAVA-Video-cap and LLAVA-Video-QA datasets we used are subsets extracted from LLaVA-Video-178K, excluding Charades, ActivityNet, and Ego4D video sources, thus preventing potential data leakage. The data we utilized includes tasks such as Temporal Sentence Grounding, Temporal Action Localization, Dense Video Captioning, VideoQA, Classification, Video Captioning, and iMOVE-IT, with the detailed composition of each task shown in Table 9.

B.4 Detailed Benchmarks and Evaluation Metrics

iMOVE is comprehensively evaluated across the following four tasks:

Temporal Video Grounding: This task aims to determine the temporal boundaries of a single event based on a text description. The datasets **Charades-STA** (Gao et al., 2017) and **ActivityNet-Captions** (Caba Heilbron et al., 2015) are used for evaluation. For this task, we report the Inter-

Task	# of Entries	Datasets
Temporal Sentence Grounding	194K	DiDeMo, HiREST, QuerYD, VTG-IT-MR, TACOS
Temporal Action Localization	45K	HACS
Dense Video Captioning	61K	COIN, ViTT, YouCook2
VideoQA	493K	EgoQA, NExT-QA, Intent-QA, CLEVRER, LLAVA-Video-QA
Classification	66K	SthSthV2, Kinetics
Video Captioning	276K	YouCook2, WebVid-stage3, LLAVA-Video-cap
iMOVE-IT	114K	self-collected

Table 9: Dataset Composition.

Task	# of Entries	Datasets
Temporal Sentence Grounding	194K	DiDeMo, HiREST, QuerYD, VTG-IT-MR, TACOS
Temporal Action Localization	45K	HACS
Dense Video Captioning	61K	COIN, ViTT, YouCook2

Table 10: Dataset Composition of the ablation study on Event-aware Spatiotemporal Efficient Modeling.

section over Union (IoU) between the predicted timestamps by the model and the ground truth annotations. Specifically, we calculate **Recall at IoU** thresholds of $\{0.3, 0.5, 0.7\}$ and their **mean IoU**.

Dense Video Captioning: This task is more complex, requiring the joint localization of key events and the generation of descriptions for each segment. The **ActivityNet-Captions** dataset is used for evaluation. We report **SODA_c** (Fujita et al., 2020), which is specifically tailored for the video's storyline, and **METEOR** (Banerjee and Lavie, 2005), which is the average of traditional METEOR scores calculated based on matched pairs between generated events and the ground truth across IoU thresholds of $\{0.3, 0.5, 0.7, 0.9\}$.

General Video Understanding: This task aims to evaluate the general short-term video understanding capabilities of iMOVE. We utilize **Video-MME** (Fu et al., 2024) and **MVBench** (Li et al., 2024), reporting their average accuracy.

Long-term Video Understanding: This task aims to evaluate the long-term video understanding capabilities of video models. We utilize the **LongVideoBench** (Wu et al., 2024) and the Long subset of **Video-MME**, reporting their average accuracy.

B.5 Explanation of Methods on ActivityNet-Captions That Are Not Strictly Zero-Shot Setting

Due to potential data leakage, some methods on ActivityNet-Captions do not strictly adhere to zero-shot settings. Below is a detailed explanation:

PiTe utilized the Video-ChatGPT (Maaz et al., 2024b) dataset in Stage 3, which was constructed using ActivityNet (Caba Heilbron et al., 2015) as the source.

Grounded-VideoLLM employed ANet-RTL, VCG-Plus-112K (Maaz et al., 2024a), Videochatgpt-100K, and Videochat2-Conv in Stage 3, all of which were constructed using ActivityNet as the source. Additionally, the TextVR (Wu et al., 2025) used in this method also utilized videos from ActivityNet.

HawkEye also utilizes VideoChatGPT and TextVR, resulting in non-strict zero-shot settings on the ActivityNet-Captions dataset.

VTimeLLM directly used the ActivityNet Captions (Krishna et al., 2017) dataset as the training set in Stage 3.

B.6 Data Composition of Ablation Study

Due to computational resource considerations, we conducted the ablation experiment for Event-aware Spatiotemporal Efficient Modeling using only a subset of the entire dataset related to temporal understanding. As shown in Table 10, this includes the tasks of Temporal Sentence Grounding, Temporal Action Localization, and Dense Video Captioning.

C Qualitative analysys

We provide a detailed qualitative comparison between iMOVE and InternVL2-4B in terms of tem-

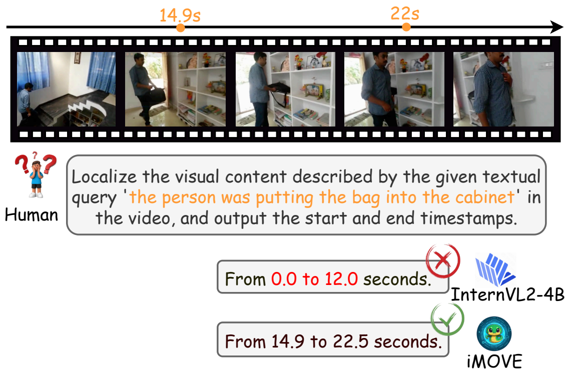

Figure 6: Qualitative comparison of the temporal grounding capabilities of iMOVE and InternVL2-4B.

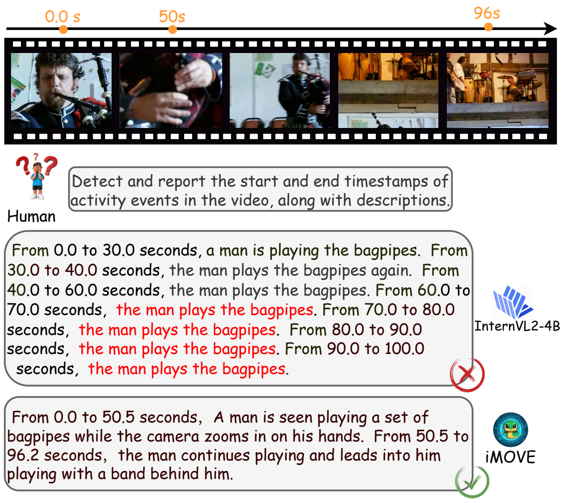

Figure 7: Qualitative comparison of the dense captioning capabilities of iMOVE and InternVL2-4B.

poral grounding, dense captioning, and long-term video understanding.

Qualitative Comparison in Temporal Grounding. As shown in Figure 6, iMOVE accurately identifies the time interval of the event "the person was putting the bag into the cabinet" from a video containing multiple events, whereas InternVL2-4B fails to pinpoint the specific video segment where the event occurs. This demonstrates iMOVE's strong fine-grained temporal perception capability.

Qualitative Comparison in Dense Captioning. As illustrated in Figure 7, iMOVE effectively captures the complete storyline of the video, accurately identifying the time intervals of various events and providing precise event descriptions. In contrast, InternVL2-4B struggles to correctly comprehend multiple events within the video, resulting in repetitive event descriptions.

Qualitative Comparison in Long-term Video Understanding. As depicted in Figure 8, thanks to its meticulously designed architecture and dataset, iMOVE accurately answers reasoning questions

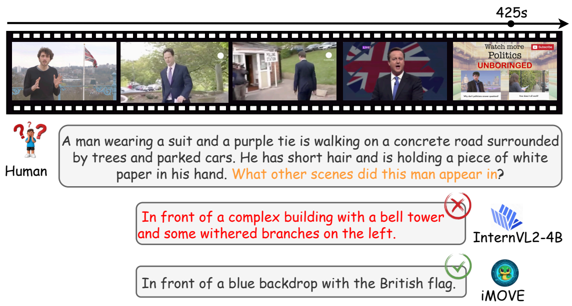

Figure 8: Qualitative comparison of the long-term video understanding capabilities of iMOVE and InternVL2-4B.

in long videos. iMOVE first locates the segments described in the questions within the long video and determines the key characteristics of the man based on the video content, using them as crucial clues when he appears in another video segment. As a short video large language model, InternVL2-4B fails to provide correct answers.

In summary, iMOVE, with its carefully designed architecture based on iMOVE-IT, adapts a short video large language model to perceive fine-grained spatiotemporal instance motions, significantly enhancing its temporal understanding, general video understanding and long-term video understanding capabilities. This improvement in general video understanding offers a substantial advantage over previous temporal Video-LLMs, which excelled in temporal understanding but lacked in general video understanding capabilities.