PiTe: Pixel-Temporal Alignment for Large Video-Language Model

Yang Liu[1,2⋆], Pengxiang Ding[1⋆], Siteng Huang[1], Min Zhang[1], Han Zhao[1], and Donglin Wang[1✉]

[1]Westlake University [2]Soochow University
{liuyang67, dingpengxiang, wangdonglin}@westlake.edu.cn

Abstract. Fueled by the Large Language Models (LLMs) wave, Large Visual-Language Models (LVLMs) have emerged as a pivotal advancement, bridging the gap between image and text. However, video making it challenging for LVLMs to perform adequately due to the complexity of the relationship between language and spatial-temporal data structure. Recent Large Video-Language Models (LVidLMs) align feature of static visual data like image into latent space of language feature, by general multi-modal tasks to leverage abilities of LLMs sufficiently. In this paper, we explore fine-grained alignment approach via object trajectory for different modalities across both spatial and temporal dimensions simultaneously. Thus, we propose a novel LVidLM by trajectory-guided **Pi**xel-**Te**mporal Alignment, dubbed **PiTe**, that exhibits promising applicable model property. To achieve fine-grained video-language alignment, we curate a multi-modal pre-training dataset PiTe-143k, the dataset provision of moving trajectories in pixel level for all individual objects, that appear and mention in the video and caption both, by our automatic annotation pipeline. Meanwhile, **PiTe** demonstrates astounding capabilities on myriad video-related multi-modal tasks through beat the state-of-the-art methods by a large margin.

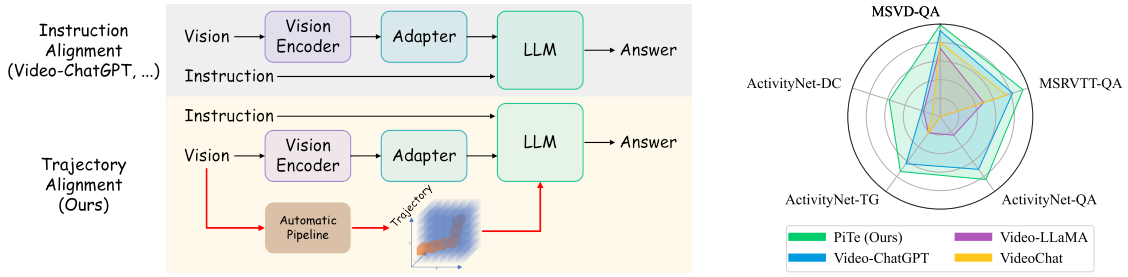
Keywords: Large Video-Language Model · Trajectory-guided Instruction Tuning · Video Understanding

1 Introduction

Large Language Models (LLMs) have rapidly gained popularity within the AI community, demonstrating astounding capabilities across a wide array of natural language tasks [4, 5, 9, 29, 37, 38]. The powerful language comprehension abilities of LLMs drive researchers to explore their utility in addressing a broader spectrum of tasks across various domains. Consequently, an increasing number of studies are focusing on developing comprehensive Large Visual-Language Models (LVLMs) to tackle vision-related tasks in zero-shot settings [6, 18, 47, 52], particularly in the video understanding [19, 23, 26, 27, 44, 46, 48]. The pursuit of generalist Large Video-Language Models (LVidLMs) will be a perennial challenge. Success in this endeavor hinges on effectively leveraging the exceptional understanding, reasoning, and generative capacities inherently present in LLMs.

One potential route towards addressing the issue is aligning visual feature into latent space of language feature. To achieve this, existing LVidLMs apply large-scale vanilla instruction tuning [19, 26, 46, 48]. However, the conventional question-answering training paradigm primarily assists LLMs in understanding visual data from a spatial perspective, posing challenges in effectively capturing

⋆ Equal contribution. ✉ Corresponding author.



(a) Illustration of training paradigm comparison. **(b)** Performance comparison.

Fig. 1: Comparison with existing LVidLMs in terms of alignment paradigm and performance. For Fig. 1b, *QA*, *TG*, *DC* denote question answering, temporal grounding and dense captioning, respectively.

temporal dynamics and spatial consistency relationships. Therefore, relying solely on instruction tuning proves insufficient for achieving comprehensive video comprehension, given the intricate spatial-temporal data structure involved. It is crucial to align different modalities across both spatial and temporal dimensions. Furthermore, offering more fine-grained cross-modal alignment guidance significantly enhances LVidLMs' ability to comprehend videos [24].

To bridge the gap, we introduce a novel LVidLM named **PiTe**, which emploies trajectories to intricately align vision and language across both spatial and temporal dimensions at the pixel level, and the distinction from conventional approaches is illustrated in Figure 1a. By requiring the model to forecast the trajectory of individual objects mentioned in the text within the video, it enables the learning of fine-grained text-to-pixel alignment through exploiting the video context along the temporal dimension and enhancing its ability to generate output based on evidence.

Subsequently, due to there are no ready-made video-language dataset with moving trajectory of objects, we curate a large-scale video-language dataset PiTe-143k through an automated annotation pipeline. Consequently, as shown in Figure 1b, the proposed **PiTe** significantly augments the LVidLM's capacity to understand videos comprehensively, leading to promising, competitive, and state-of-the-art performance in question-answering, temporal grounding, and dense captioning tasks under zero-shot conditions.

Overall, our principal contributions in this paper are summarized as follows:

• We curate a large scale video-language dataset PiTe-143k with trajectory for all individual objects by automatic annotation pipeline.

• We propose a novel LVidLM **PiTe** that utilize trajectory to align video and language features across both spatial and temporal dimensions.

• Extensive experimental results and analysis on myriad datasets for zero-shot video question answering, temporal grounding, and dense captioning tasks demonstrate the superiority of **PiTe**.

2 Related Work

2.1 Large Language Models

Over the last few years, pioneering foundation language models like GPT-1 [32], BERT [7], GPT-2 [33], and T5 [34] laid the groundwork, but GPT-3 [4] groundbreaking model parameters to 175 billion size to achieve remarkable zero-shot performance. Besides, research on scaling law [14] has steered language models to a larger scale. Therefore, driven by the success of InstructGPT [29]

and ChatGPT [28] which training by reinforcement learning with human feedback (RLHF) based on GPT-3, Large Language Models (LLMs) has made waves in the natural language processing (NLP) community due to its capabilities in language understanding, logical reasoning, and generation. The GPT's success suggest a promising path towards building LLMs, several open-source LLMs have been proposed following it with similar performance including OPT [49], BLOOM [36], GLM [9], LLaMA [37, 38], and Vicuna [5]. Our investigation delves into leveraging the striking language comprehension and zero-shot generalization abilities of LLMs beyond the confines of linguistic modalities. Specifically, we aim to extend these capabilities to multi-modal scenarios, thereby exploring their potential in processing diverse forms of information across different modalities.

2.2 Large Visual-Language Models

The surge of LLMs has lead to major advancements in NLP tasks, and also has incited interest in developing Large Visual-Language Models (LVLMs). Building a unify LLM with visual inputs for visual language tasks thus remains one of the most important desiderata for LVLMs. Flamingo [1] and OpenFlamingo [3] fuse visual information into intermediate embedding for a frozen LLM by cross-attention mechanism, and train on billions of image-text pairs to align visual and linguistic modalities. Similarly, BLIP-2 [18] introduced the concept of Q-Former to align visual features more effectively with language space. Moreover, MiniGPT-4 [52] enhances its usability significantly by further fine-tuning on more detailed image descriptions with just one projection layer to align a frozen visual encoder with a frozen LLM, and the LLaVA series [21, 22] use simply a multi-layer perception (MLP) in place of the Q-Former and two-stage instruction tuning to enhance this process. Furthermore, PixelLLM [42] leverage the location coordinate of every word in the caption in the image as the connection between different modalities to strengthen the model's performance for the object detection task. Our primary focus lies in transferring the exceptional language comprehension capabilities of LLMs to the analysis of dynamic, continuous visual data found in videos, as opposed to static visual data such as images.

2.3 Large Video-Language Models

Recently, many efforts have been made to transfer the task-handling capability of LVLMs to the video modality, leading to the emergence of Large Video-Language Models (LVidLMs) like VideoChat [19], Video-LLaMA [46], and Video-ChatGPT [26]. Prior researches have demonstrated the capability of LLMs to perform diverse tasks on video content, guided by user instructions through a two-stage training process. These studies align static visual features with LLMs, followed by instruction tuning on datasets annotated either by GPT or humans. Despite being effective in video understanding, the lack of fine-grained spatial-temporal modeling in these models prevents them from understanding or locating object in detail or specific segments. We propose a novel fine-grained alignment strategy at the pixel level across spatial and temporal dimensions to enhance the ability of LLMs to comprehensively analyze video content, thereby facilitating a more detailed understanding of the visual information presented.

3 PiTe-143k Dataset

To facilitate fine-grained multi-modal alignment research at the pixel level, we introduce a large-scale video-language dataset PiTe-143k. This dataset fills a crucial gap in the existing resources by

Table 1: Comparison between PiTe-143k and existing video instruction datasets.

Dataset	Total Dur.	Avg. Dur.	#Videos	#Events	Temporal Localization	#Objects Trajectories
VideoChat [19]	41h	18s	8.2k	✗	✗	✗
Valley [25]	608h	40s	54.7k	✗	✗	✗
Video-ChatGPT [26]	432h	**117s**	13.3k	✗	✗	✗
PiTe-143k (Ours)	**2086.44h**	52.18s	**143.64k**	**343.93k**	✓	1.02M

providing extensive object moving trajectories with video instruction, which were previously unavailable in ready-made datasets. PiTe-143k constructed based on InternVid-10M-FLT [13,40] that each instance contains and entire video and multiple clip captions with start-stop timestamps. As shown in Table 1, PiTe-143k comprises 343.93 thousand event segments and 1.02 million moving trajectory for all individual objects that appear in both visual and textual modalities. To facilitate this objective, we establish an automatic annotation pipeline for PiTe-143k, fostering the advancement of LVidLMs for nuanced pixel-level video comprehension.

The automatic annotation pipeline for PiTe-143k comprises two primary stages, as depicted in Fig. 2: (1) Stage 1 involves the noun phrases extraction and referring expressions segmentation, thereby generating object masks within the frame for all individual objects referenced in the event caption; (2) Stage 2 centers on point tracking to capture the moving trajectories corresponding to the masks obtained in Stage 1.

3.1 Referring Expression Segmentation

In stage 1, we aim to build closely fine-grained connection between video and language. To this end, we extract all noun phrases from caption and find the corresponding objects in the clip.

At inception, we leverage constituency parser SuPar [50,51] for language to extract noun phrase as shown in Fig. 3. Notably, in order to pass the simplest and most straightforward language instructions in next step, we only extract noun phrase from the lowest layer. For example in Fig. 3a, we consider two noun phrases *a pen* and *a white table*, but the parent node of the former that denotes *a pen on a white table* not in our consideration because of the complexity of its composition. Following this, we utilize GLaMM [35], the first LVLM that can generate natural language responses seamlessly intertwined with corresponding object segmentation masks, to obtain the corresponding segmentation mask in the first frame of the clip for the text-based referring expression. While certain objects in the video, such as *a pen* as illustrated in Fig.3a and Fig.2, may be too small to accurate detection. In such challenging cases, we disregard the trajectory information of the noun phrase. Despite this limitation, it has a minimal impact on the overall performance when utilizing extensive pre-training data on a large scale. Meanwhile, leveraging the exceptional language comprehension capabilities of LLMs, GLaMM can effectively filter out invalid referring expressions, those that do not constitute legal object references, such as *front* as depicted in Fig. 3b.

3.2 Point Tracking

In stage 2, we aim to transfer the connection constructed in the previous stage to video, expanding out the temporal dimension specific to video compared to image. To this end, we track all individual

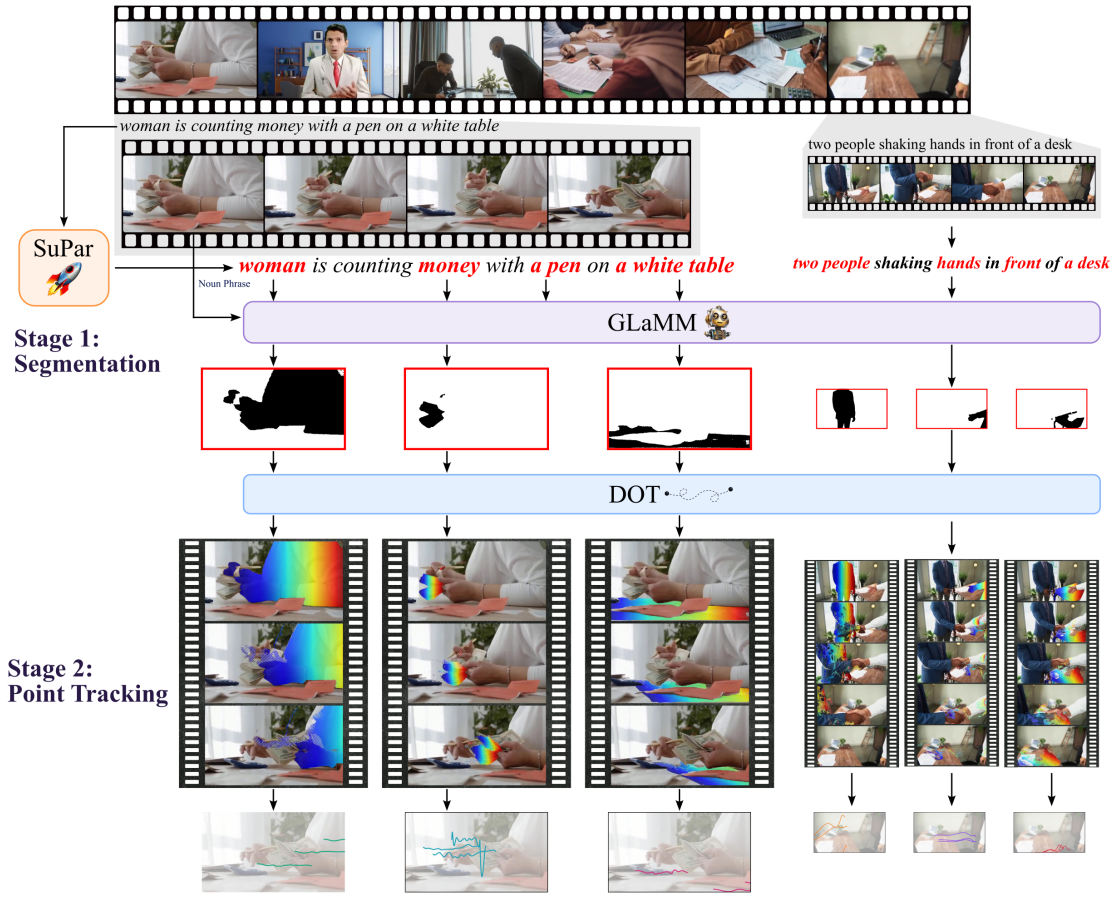
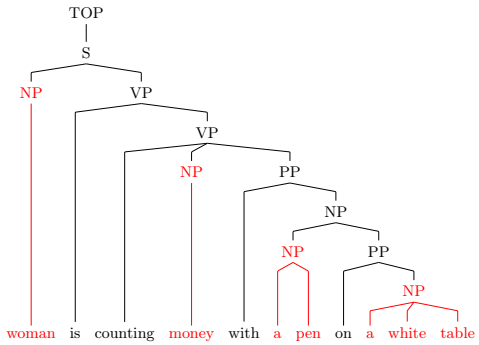


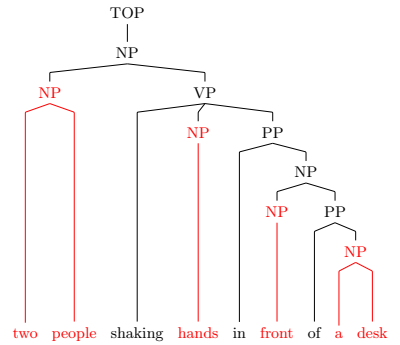
Fig. 2: Automatic annotation pipeline for PiTe-143k. The video sample in the figure showcases two events positioned at the commencement and conclusion of the video. The procedure for extracting noun phrases by SuPar [50, 51] is elucidated in Fig. 3.

objects in their clip to obtain the trajectory, the trajectory indicates the connection between video and language in both spatial and temporal dimensions.

The stage 2 commences when we employ DOT [17], a simple-yet-efficient method for tracking point to recover the trajectory of any scene point, for each clip to capture the trajectories for any point in first frame. According to our observation, the caption of each clip mainly describes simple video content in short sentences, so most of the caption corresponds to just one scene clip, which enables us to track objects that identified in the first frame. Subsequently, filter trajectories according to the segmentation mask of objects obtained in stage 1. So far, we obtain the trajectories for all objects in each clip for each video, we create the connection between video and language from both spatial and temporal through the trajectories, the existence of trajectories in video denotes whether the object exist, and the value of trajectory represent where the object exist in video. Lastly, we utilize the k-means++ [2] clustering algorithm to condense trajectories into three key points, effectively reducing computational demands. This approach is founded on the premise that



(a) The extraction of the NP for the first event caption of the sample video depicted in Fig. 2.



(b) Referring to Fig. 2, the extraction of the NP for the second event caption.

Fig. 3: Two samples of constituency parser for Noun Phrase (NP) extraction.

three points adequately capture the typical geometric shape of objects, striking a balance between precision and computational efficiency. Furthermore, we conduct a comparative analysis of the performance using various key tracking points, as discussed in Section 5.3.

4 PiTe

In this section, we propose a novel Large Video-Language Model (LVidLM), **PiTe**, which align video and language by trajectories across both spatial and temporal dimension. Fig. 4 illustrates an overview of **PiTe**.

4.1 Architecture

PiTe is composed of a vision encoder to encode frames from video implemented as ViT [8], a vision adapter to project visual feature to semantic space of LLMs implemented as a linear projection layer, a LLM Vicuna v1.5 [5], and a localization projector or trajectory projector in separate training stage to guide LLMs to understand visual information implemented as a linear projection.

Vision Encoder. Raw video data can be expressed as multiple frames such as $\mathbf{v} = \{f_1, f_2, \ldots, f_N\} \in \mathbb{R}^{N\times H\times W\times C}$ (frames$\times$ height $\times$ width $\times$ channels). Following previous studies [22,26,46], we adopt ViT-L/14 [8] pre-trained from CLIP [31] as the vision encoder ViT to encode visual data. We uniformly sample N frames for video $\mathbf{v}$, and encode i-th frame f_i through the vision encoder ViT:

$$\left\{v_i^{cls}, v_i^1, v_i^2, \ldots, v_i^P\right\} = \text{ViT}\left(f_i\right), \tag{1}$$

where P denotes the number of patches in the vision encoder ViT.

Visual Adapter. A simpler projector forces the LLMs to learn more on handling visual inputs, leading to better generalization [20]. Hence, we utilize the global feature v_i^{cls} from vision encoder ViT as the representation for the i-th frame f_i, and we apply a linear projection layer $\mu(\cdot)$ to connect the frame feature into the word embedding space of LLMs:

$$z_i = \mu\left(v_i^{cls}\right). \tag{2}$$

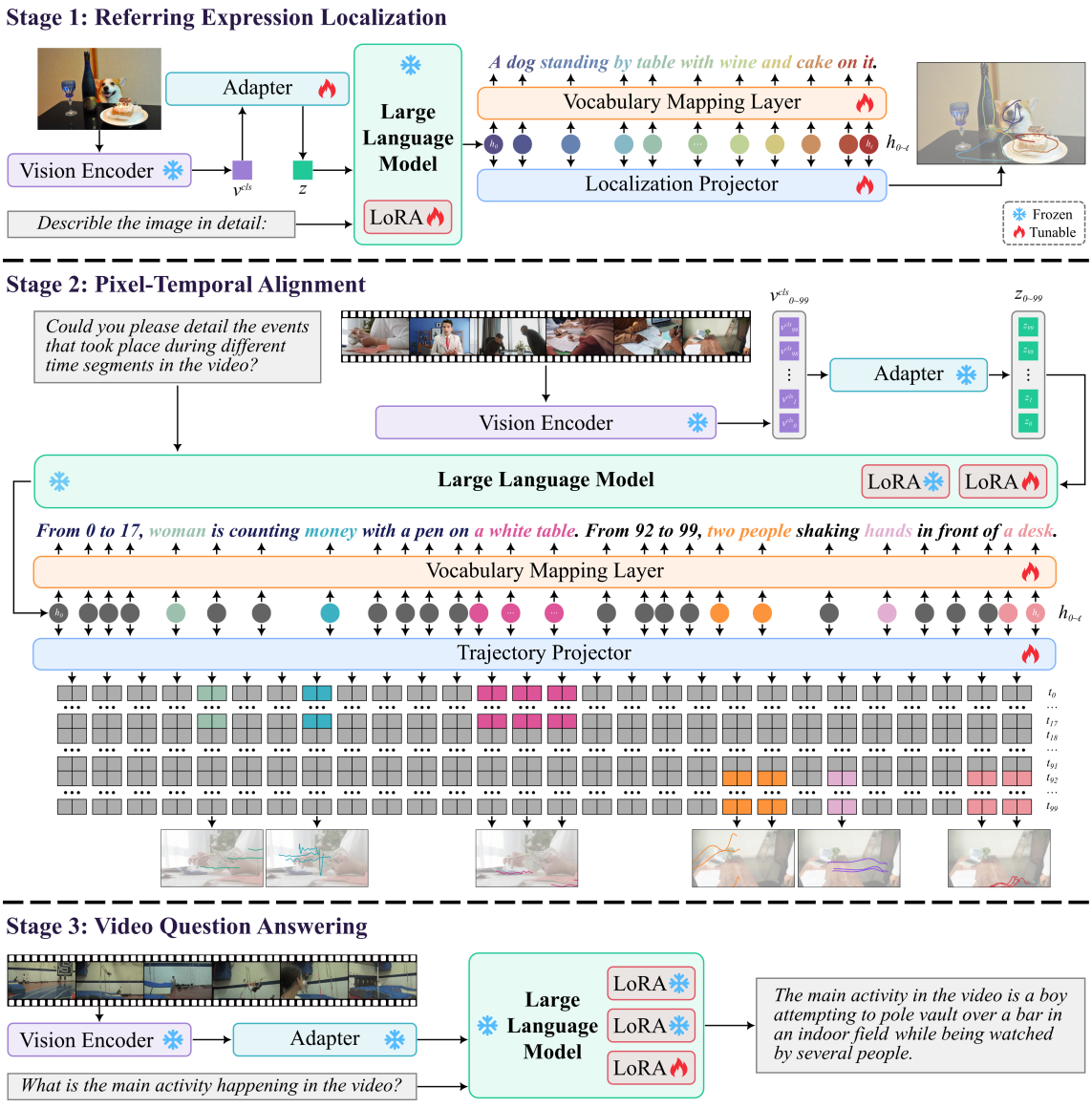


Fig. 4: Schematic of **PiTe** framework for video-language alignment.

Subsequently, a sequence of frame tokens $\mathbf{z} = \{z_1, z_2, \ldots, z_N\} \in \mathbb{R}^{N \times d}$ becomes the input that LLMs can understand, d denotes the hidden dimension of LLMs.

Large Language Model. After we tokenize and encode video into frame tokens $\mathbf{z}$, we concatenate it with textual tokens $\mathbf{w} = \{w_1, w_2, \ldots, w_L\} \in \mathbb{R}^L$ and feed as the input to LLMs, we treat visual input as a foreign language in this process. Based on this, LLMs can further encode the input

sequence to understand the video and text, then reasoning and generation the response using auto-regressive decoding as follows:

$$h_i = \mathrm{LLM}^{-}\left(\mathbf{z}, \mathbf{w}_{1:i-1}\right), \tag{3}$$

$$w_i = \operatorname{argmax}\left(\mathbf{m}_v \cdot h_i\right), \tag{4}$$

where LLM^{-} denotes the LLM without the last vocabulary mapping layer, h_i denotes the hidden states of i-th token generated by LLM^{-}, $\mathbf{m}_v \in \mathbb{R}^{|V| \times d}$ denotes the weight of the linear vocabulary mapping layer.

4.2 Training Strategy

For **PiTe** model training, we consider a three stage instruction-tuning procedure, as depicted in Fig. 4: (1) Stage 1 centered around training adapters using image-caption pairs; (2) Stage 2 is focuses on aligning video and language features through trajectories; (3) Stage 3 is dedicated to enhancing the model's comprehension by following human instructions through high-quality dialogue instruction tuning.

Stage 1: Referring Expression Localization. At the initial stage, we aim to train the visual adapter that can align visual features with semantic space of LLMs. To this end, we employ Localized Narratives dataset [30] that contains annotations of human annotators narrating a given image, together with a mouse trajectory of the annotators' attention during the narration. This gives synchronized locations for all words in the narration sentence, the cross-modal attention of human can be used to train our model as condition to bridge vision and language.

There is only one visual tokens $\mathbf{z} = \{z_1\} \in \mathbb{R}^{1 \times d}$ in this training stage for image instead of video, this can align vision with language in spatial to train adapter without consider temporal information. To use the same language features for localization, we simply add a multi-layer perception (MLP) as localization projector $\varphi(\cdot)$ parallel with the vocabulary mapping layer, which maps the language feature to a 2-dimension location:

$$p_i = \varphi(h_i), \tag{5}$$

where p_i denotes the predicted coordinate for textual token w_i.

Overall, the leaning objective in the first stage was calculated by standard label-smoothed cross-entropy loss to train captioning output, and L_1 regression loss to train the localization output:

$$\mathcal{L}_1 = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\mathrm{CE}\left(\mathrm{LLM}\left(\mathbf{z}, \mathbf{w}_{1:i-1}\right), w_i\right) + \lambda \mid \hat{p}_i - p_i \mid\right), \tag{6}$$

where ℓ is the length of the generated sequence, $\hat{p}_i$ represents the ground truth of location, and $\mathrm{CE}(\cdot, \cdot)$ denotes the cross-entropy function. To enhance training efficiency, we utilize LoRA [12] for fine-tuning the LLM.

Stage 2: Pixel-Temporal Alignment. After stage 1, the LLM model becomes proficient in understanding visual information. In the stage 2, we aim to train the LLM to understand sequential frames in video. To achieve this target, we curate a detailed object tracking dataset PiTe-143k, as described in Section 3 that uses trajectory as condition to bridge vision and language across both spatial and temporal dimensions. Therefore, the alignment guidance in pixel level improves the model's video fine-grained understanding reliability and overall usability.

Table 2: Training hyper-parameters of PiTe.

Configuration	Stage 1	Stage 2	Stage 3	Configuration	Stage 1	Stage 2	Stage 3
Vision Encoder	OpenAI-CLIP-L/14			Learning Rate	0.0001		
Image/Frame Resolution	224×224			LoRA	r=64 & α=128		
Adapter Parameter	Tunable	Frozen	Frozen	Numerical Precision	BFloat16		
Video Frames	100			Epoch	1	2	2
LLM	Vicuna-7B/13B-v1.5			Global Batch Size	256		
LLM Sequence Length	2048			Learning Rate Schedule	Cosine Decay		
Optimizer	AdamW			Warm-up Ratio	0.03		

Similar to stage 1, we use the same language features for alignment by a MLP as trajectory projector $\rho(\cdot)$ to map the language feature to a 2-dimension location:

$$\mathbf{p}_i = \rho(h_i), \tag{7}$$

where $\mathbf{p}_i$ denotes the trajectory matrix in P points and N frames for textual token w_i. Here, we define p_{ijk} indicates the coordinate for token w_i for i-th point for model tracking in frame f_j.

Overall, the leaning objective in stage 2 was calculated by standard label-smoothed cross-entropy loss to train the generation output, and L_1 regression loss to train the trajectory output as the condition:

$$\mathcal{L}_2 = \frac{1}{\ell}\sum_{i=1}^{\ell}\left(\mathrm{CE}\left(\mathrm{LLM}\left(\mathbf{z}, \mathbf{w}_{1:i-1}\right), w_i\right) + \frac{\lambda}{P \cdot N}\sum_{j=1}^{P}\sum_{k=1}^{N} \mid \hat{p}_{ijk} - p_{ijk} \mid \right), \tag{8}$$

where P is the number of the points for model tracking to generate trajectory, and $\mathbf{z} \in \mathbb{R}^{N\times d}$ represents the sequence of visual embedding. We merge the LoRA trained in the stage 1 with the original model and introduce a new LoRA module.

It is notable that we use localization projector $\varphi(\cdot)$ trained in previous stage to initialize the trajectory projector $\rho(\cdot)$. Specifically, we define the weight of localization projector $\varphi(\cdot)$ and trajectory projector $\rho(\cdot)$ as $\mathbf{m}_\varphi \in \mathbb{R}^{P\cdot N\cdot 2\times d}$ and $\mathbf{m}_\rho \in \mathbb{R}^{2\times d}$, respectively. localization projector $\varphi(\cdot)$ maps a 2-dimension coordinate on the input image for each token of the LLM output, as for trajectory projector $\rho(\cdot)$, it also output 2-dimension coordinates, but more than $P \cdot N$ times for P points to tracking for N frames. For each point of each frame, the parameter of trajectory projector $\rho(\cdot)$ initialized by the localization projector $\varphi(\cdot)$:

$$\mathbf{m}_\varphi = \overbrace{\mathbf{m}_\rho \oplus \mathbf{m}_\rho \oplus \cdots \oplus \mathbf{m}_\rho}^{P\cdot N}, \tag{9}$$

where $\oplus$ denotes concatenation of matrix in first dimension.

Beyond trajectories, our model is attuned to temporal boundaries within the generated text. Specifically, we structure the generation as `..., from s to e` or `From s to e, ...` to facilitate the model to learn in temporal dimension. Here, `...` encapsulates the event description, while `s` and `e` denote the frame indexes corresponding to the start and end timestamps of event, respectively. This approach further augments the model's understanding of temporal boundaries [13].

Dissimilar to the initial training stage, not all generated words are associated with trajectories. In cases where objects lack trajectories or vanish from view over time, we uniformly assign the coordinates of their ground truth as $(-1, -1)$ to signify their absence.

Table 3: Comparison between different LVidLMs on zero-shot question-answer.

Model	LLM Size	MSVD-QA [41]		MSRVTT-QA [43]		ActivityNet-QA [45]	
		Accuracy↑	Score↑	Accuracy↑	Score↑	Accuracy↑	Score↑
FrozenBiLM [44]	1B	32.2	-	16.8	-	24.7	-
LLaMA-Adapter [48]	7B	54.9	3.1	43.8	2.7	34.2	2.7
VideoChat [19]	7B	56.3	2.8	45.0	2.5	-	2.2
Video-LLaMA [46]	7B	51.6	2.5	29.6	1.8	12.4	1.1
Video-ChatGPT [26]	7B	64.9	3.3	49.3	2.8	35.2	2.7
PG-Video-LLaVA [27]	7B	64.1	3.7	51.6	3.3	39.9	**3.3**
PiTe (Ours)	7B	**68.4**	**3.9**	**56.4**	**3.5**	**42.0**	**3.3**
PiTe (Ours)	13B	**71.6**	**4.0**	**57.7**	**3.5**	**42.2**	**3.4**

Stage 3: Video Question Answering. Following stage 2, we incorporate high-quality dialogue data Valley [25] and Video-ChatGPT [26] in one turn for instruction tuning, enabling the model to follow human instructions for more accurate and generalize capabilities of video understanding.

The leaning objective in third stage was calculated by standard label-smoothed cross-entropy loss for auto-regression generation:

$$\mathcal{L}_3 = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathrm{CE}\left(\mathrm{LLM}\left(\mathbf{z}, \mathbf{w}_{1:i-1}\right), w_i\right). \tag{10}$$

Similar to stage 2, we merge the LoRA trained in the stage 1 and stage 2 with the original model and introduce a new LoRA module.

5 Experiments

5.1 Experimental Setup

Tasks, Datasets, and Evaluation Metrics. We conduct a quantitative evaluation of LVidLMs' video understanding capabilities across three tasks: (1) Video Question Answering: This task assesses the comprehensive video comprehension abilities of LVidLMs by requiring the model to answer a variety of questions about the video content based on its understanding. We perform this task on three datasets: MSVD-QA [41], MSRVTT-QA [43], and ActivityNet-QA [45]. The evaluation pipeline for video understanding follows Video-ChatGPT [26], and we report the accuracy and score, which is assessed using GPT-Assistant [28]. (2) Video Temporal Grounding: This task evaluates LVidLMs' capacity to discern the starting and ending timestamps of a segment corresponding to the description of a video clip. This task demands the model to effectively grasp the temporal aspects of the video. We conduct this task on the ActivityNet Captions dataset [15] and calculate Intersection over Union (IoU) between the model-generate time segments and the ground truth time segments. We report mean IoU (mIoU) and Recall@1, IoU$\geq m$ (R@m) metric, where m values are set at $\{0.3, 0.5, 0.7\}$. (3) Video Dense Captioning: This task requires the model to produce all events depict in the video along with their corresponding start and end timestamps. It necessitates the model to comprehend both the spatial and temporal dimensions of the video simultaneously.

Table 4: Comparison between different LVidLMs in temporal video grounding and dense video captioning tasks on ActivityNet [11].

Model	LLM Size	Temporal Grounding				Dense Captioning		
		R@0.3↑	R@0.5↑	R@0.7↑	mIoU↑	SODA_c↑	CIDEr↑	METEOR↑
VideoChat [19]	7B	8.8	3.7	1.5	7.2	0.9	2.2	0.9
Video-LLaMA [46]	7B	6.9	2.1	0.8	6.5	1.9	5.8	1.9
Video-ChatGPT [26]	7B	26.4	13.6	6.1	18.9	1.9	5.8	2.1
PiTe (Ours)	7B	**30.4**	**17.8**	**7.8**	**22.0**	**5.1**	**21.7**	**5.8**
PiTe (Ours)	13B	**37.2**	**23.7**	**10.9**	**26.0**	**5.9**	**26.5**	**6.6**

We conduct this task on the ActivityNet Captions dataset [15]. Initially, we reported SODA_c [10], followed by averages of CIDEr [39] and METEOR [16] under different IoU thresholds of 0.3, 0.5, 0.7, 0.9 based on generate events and ground truth matched pairs to provide a comprehensive analysis. In this paper, all experiments were conducted in a zero-shot setting, and higher values of all evaluation metrics indicate superior performance.

Implementation Details. In this paper, we employ Vicuna v1.5 [5] as LLM to train the **PiTe** model at two scales: 7B and 13B. Leveraging the efficiency of LoRA [12], the training of the 7B model can be completed in approximately 10 hours using a single Nvidia 8-A100 (80GB VRAM) node, while the 13B model requires around 17 hours. More hyper-parameter settings are shwon in Table 2.

5.2 Main Result

Table 3 and 4 present the comparative performance of the **PiTe** against state-of-the-art baselines on myriad video understanding datasets .

Question Answering. As illustrated in Table 3, **PiTe** consistently outperforms the state-of-the-art pure instruction-tuning baselines in terms of all metrics on all datasets. Compared to the top-performing baselines in each dataset, **PiTe** exhibited notable improvements in the average question answering accuracy, achieving a maximum enhancement of 4.8 and an average improvement of 3.7. For example, **PiTe** substantially improves accuracy by 64.9 to 68.4 compared to Video-ChatGPT [26] in MSVD-QA dataset [41]. The results showcasing **PiTe**'s proficiency in video comprehension and its capacity to deliver contextually relevant responses according to the given instructions.

Temporal Grounding. As depicted in Table 4, **PiTe** achieves state-of-the-art performance in the video temporal grounding task across all metrics as well, demonstrating improvements ranging from 18.9 to 22.0 in mIoU compared to Video-ChatGPT [26]. This clearly indicates that trajectory alignment greatly enhances the ability of capture events in temporal dimension for LVidLMs. The incorporation of object trajectories in the temporal dimension of the trajectory matrix equips the model with a precise understanding of temporal event boundaries, thereby establishing a solid foundation for accurate event localization.

Table 5: Ablation study of the three-stage training strategy.

Method	MSVD-QA		ActivityNet						
			Temporal Grounding				Dense Captioning		
	Accuracy	Score	R@0.3	R@0.5	R@0.7	mIoU	SODA_c	CIDEr	METEOR
PiTe (Ours)	**68.4**	**3.9**	**30.4**	**17.8**	**7.8**	**22.0**	**5.1**	**21.7**	**5.8**
w/o initialize	68.2	**3.9**	22.8	10.5	4.6	17.1	**5.1**	**21.7**	**5.8**
w/o trajectory	68.1	**3.9**	23.9	12.8	5.7	17.4	5.0	21.4	**5.8**

Dense Captioning. The outcomes of the dense captioning task, as delineated in Table 4, reveal that **PiTe** consistent boost compared to all state-of-the-art baselines. Particularly noteworthy is the substantial 15.9 increase in the CIDEr metric [39] when compared to Video-ChatGPT [26]. This underscores the significance of fine-grained alignment in both spatial and temporal dimensions through trajectories, implying that **PiTe** acquires more generalized and detailed representations to offer more sophisticated event descriptions and accurate event temporal boundaries.

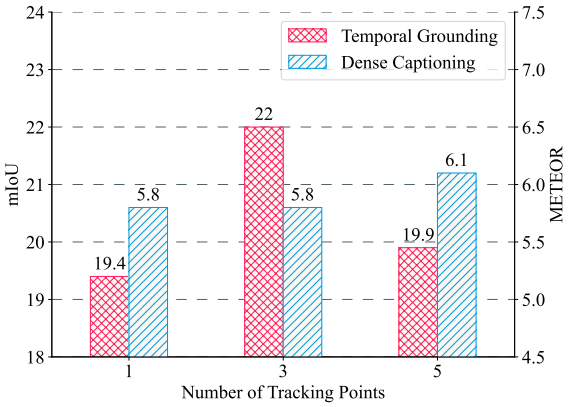
5.3 Analysis

Ablation Study. As reported in Table 5, we conduct ablation experiments on MVSD-QA [41] for question answering and ActivityNet Captions [15] for temporal grounding to verify the individual effects of the proposed contributions under the following settings: (1) w/o initialize: we remove the initialization strategy that use weight of localization projector to initialize trajectory projector; (2) w/o trajectory: we abandon the fine-grained alignment strategy via trajectory.

From the experimental results in Table 5, it can be observed that: (1) Eliminating the initialization strategy for the trajectory projector in **PiTe** reduces the model's reasoning capabilities and temporal boundary awareness. However, the performance in dense captioning generation remains consistent. This observation suggests that the model maintains its basic ability in comprehending visual content under trajectory-guided training. (2) The removal of the trajectory-guided training diminishes almost all the capability of **PiTe**, including dense captioning. (3) Without trajectory-guided training, **PiTe** demonstrates superior performance compared to trajectory-guided training without the initialization strategy for the trajectory projector in temporal grounding. This outcome highlights the difficulty of trajectory-guided training without initialization from a pre-trained localization projector, as the instability of parameters can impede the model's perception to accurately perceive visual temporal information.

Exhibition. To better illustrate the video dialogue performance of **PiTe**, we present a qualitative example, as shown in Fig. 5a. The illustration from the upper portion of the figure demonstrates **PiTe**'s capability not only to provide precise responses to instruction queries but also to enhance the output with more detailed and accurate video information. The example in the lower segment of the figure highlights the model's proficiency in understanding instruction and capturing event, enabling precise delineation of temporal boundaries within the video, despite the constraint of a 100-frame sampling limit.

(a) Examples of PiTe's video understanding capabilities.



(b) Performance comparison for different tracking point quantity.

Fig. 5: PiTe's video understanding capabilities and performance comparison across varying tracking point quantities.

Impact of Tracking Point Quantity. In Fig. 5b, we vary the tracking point quantity P in set of $\{1, 3, 5\}$. The efficacy of dense captioning tasks tends to improve with an increase in tracking points. However, it is observed that the temporal grounding task undergoes an initial substantial improvement, only to be succeeded by a rapid decline. Less number of tracking points fails to accurately capture the object's geometry, thereby hindering the pixel-level cross-modal alignment guidance for the model. Conversely, a higher quantity of points can enhance the model's comprehension of pure visual information; however, it also introduces noise to make training more challenging. Overall, that the optimal value of P may be different for different tasks, we set $P = 3$ due to its performance maintain stability over multiple tasks.

6 Conclusion

In this paper, we focus on enhancing the performance of Large Video-Language Models (LVidLMs) by incorporating trajectory-based alignment across different modalities. To achieve fine-grained alignment between video and language across spatial and temporal dimensions, we initially curate a comprehensive multi-modal object tracking dataset, PiTe-143k, using a fully automated annotation pipeline. This dataset was developed to address the lack of large-scale video-language datasets that include multi-object moving trajectories. Subsequently, we introduce a novel **Pi**xel-**Te**mporal (**PiTe**) alignment strategy that leverages trajectory-guided pre-training to address the inherent challenges faced by LVidLMs. Through comparative analyses, we evaluate **PiTe** against state-of-the-art models and competitive baselines across various tasks in a zero-shot setting, including question-answering, temporal grounding, and dense captioning, showcasing the superior performance of **PiTe** with more sophisticated event descriptions and accurate event temporal boundaries.

Acknowledgments

This work was supported by the National Science and Technology Innovation 2030 - Major Project (Grant No. 2022ZD0208800), and NSFC General Program (Grant No. 62176215).

References

1. Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Proc. of NeurIPS (2022)
2. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proc. of SODA. pp. 1027–1035 (2007)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S.Y., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. CoRR (2023)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proc. of NeurIPS (2020)
5. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (2023)
6. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Proc. of NeurIPS (2023)
7. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proc. of AACL. pp. 4171–4186 (2019)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. of ICLR (2021)
9. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: GLM: general language model pretraining with autoregressive blank infilling. In: Proc. of ACL. pp. 320–335 (2022)
10. Fujita, S., Hirao, T., Kamigaito, H., Okumura, M., Nagata, M.: SODA: story oriented dense video captioning evaluation framework. In: Proc. of ECCV. pp. 517–531 (2020)
11. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proc. of CVPR. pp. 961–970 (2015)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: Proc. of ICLR (2022)
13. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: Vtimellm: Empower llm to grasp video moments. In: Proc. of CVPR. pp. 14271–14280 (2024)
14. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. CoRR (2020)
15. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: Proc. of ICCV. pp. 706–715 (2017)
16. Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proc. of WMT@ACL. pp. 228–231 (2007)
17. Le Moing, G., Ponce, J., Schmid, C.: Dense optical tracking: Connecting the dots. In: Proc. of CVPR. pp. 19187–19197 (2024)
18. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. of ICML. pp. 19730–19742 (2023)
19. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. CoRR (2023)
20. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proc. of CVPR. pp. 26689–26699 (2024)

21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proc. of CVPR. pp. 26296–26306 (2024)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proc. of NeurIPS (2023)
23. Liu, Y., Ping, H., Zhang, D., Sun, Q., Li, S., Zhou, G.: Comment-aware multi-modal heterogeneous pre-training for humor detection in short-form videos. In: Proc. of ECAI. pp. 1568–1575 (2023)
24. Liu, Y., Shen, T., Zhang, D., Sun, Q., Li, S., Zhou, G.: Comment-aided video-language alignment via contrastive pre-training for short-form video humor detection. In: Proc. of ICMR. pp. 442–450 (2024)
25. Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. CoRR (2023)
26. Maaz, M., Rasheed, H.A., Khan, S.H., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2024)
27. Munasinghe, S., Thushara, R., Maaz, M., Rasheed, H.A., Khan, S., Shah, M., Khan, F.S.: Pg-video-llava: Pixel grounding large video-language models. CoRR (2023)
28. OpenAI: Chatgpt: Optimizing language models for dialogue (2023)
29. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Proc. of NeurIPS (2022)
30. Pont-Tuset, J., Uijlings, J.R.R., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: Proc. of ECCV. pp. 647–664 (2020)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proc. of ICML. pp. 8748–8763 (2021)
32. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training (2018)
33. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners (2019)
34. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. pp. 140:1–140:67 (2020)
35. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proc. of CVPR. pp. 13009–13018 (2024)
36. Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., Gallé, M., Tow, J., Rush, A.M., Biderman, S., Webson, A., Ammanamanchi, P.S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A.V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P.O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A.F., Alfassy, A., Rogers, A., Nitzav, A.K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D.I., et al.: BLOOM: A 176b-parameter open-access multilingual language model. CoRR (2022)
37. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR (2023)
38. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. CoRR (2023)

39. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proc. of CVPR. pp. 4566–4575 (2015)
40. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., Qiao, Y.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: Proc. of ICLR (2024)
41. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017. pp. 1645–1653 (2017)
42. Xu, J., Zhou, X., Yan, S., Gu, X., Arnab, A., Sun, C., Wang, X., Schmid, C.: Pixel-aligned language model. In: Proc. of CVPR. pp. 13030–13039 (2024)
43. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Proc. of CVPR. pp. 5288–5296 (2016)
44. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. In: Proc. of NeurIPS (2022)
45. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proc. of AAAI. pp. 9127–9134 (2019)
46. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proc. of EMNLP. pp. 543–553 (2023)
47. Zhang, M., Huang, S., Li, W., Wang, D.: Tree structure-aware few-shot image classification via hierarchical aggregation. In: Proc. of ECCV. pp. 453–470 (2022)
48. Zhang, R., Han, J., Liu, C., Zhou, A., Lu, P., Qiao, Y., Li, H., Gao, P.: LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In: Proc. of ICLR (2024)
49. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M.T., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: OPT: open pre-trained transformer language models. CoRR (2022)
50. Zhang, Y., Li, Z., Zhang, M.: Efficient second-order treecrf for neural dependency parsing. In: Proc. of ACL. pp. 3295–3305 (2020)
51. Zhang, Y., Zhou, H., Li, Z.: Fast and accurate neural CRF constituency parsing. In: Proc. of IJCAI. pp. 4046–4053 (2020)
52. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: Proc. of ICLR (2024)