

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin¹, Yang Ye¹, Bin Zhu¹, Jiaxi Cui⁴,
Munang Ning^{1,2,3}, Peng Jin^{1,2,3}, Li Yuan^{1,2,3}

¹Peking University Shenzhen Graduate School, ²Peng Cheng Laboratory,
³AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School,
⁴PandaVilla Tech Limited

Correspondence: yuanli-ece@pku.edu.cn

GitHub: <https://github.com/PKU-YuanGroup/Video-LLaVA>

Abstract

Large Vision-Language Model (LVLM) has enhanced the performance of various downstream tasks in visual-language understanding. Most existing approaches encode images and videos into separate feature spaces, which are then fed as inputs to large language models. However, due to the lack of unified tokenization for images and videos, namely misalignment before projection, it becomes challenging for a Large Language Model (LLM) to learn multi-modal interactions from several poor projection layers. In this work, we unify visual representation into the language feature space to advance the foundational LLM towards a unified LVLM. As a result, we establish a simple but robust LVLM baseline, **Video-LLaVA**, which learns from a mixed dataset of images and videos, mutually enhancing each other. As a result, Video-LLaVA outperforms Video-ChatGPT by **5.8%, 9.9%, 18.6%, and 10.1%** on MSRVTT, MSVD, TGIF, and ActivityNet, respectively. Additionally, our Video-LLaVA also achieves superior performances on a broad range of 9 image benchmarks. Notably, extensive experiments demonstrate that Video-LLaVA mutually benefits images and videos within a unified visual representation, outperforming models designed specifically for images or videos. We aim for this work to provide modest insights into the multi-modal inputs for the LLM.

1 Introduction

Recently, LLMs have gained rapid popularity in the AI community, such as GPT-3.5, GPT-4 (OpenAI, 2023), PaLM (Bi et al., 2020; Anil et al., 2023), and BLOOM (Scao et al., 2022). They rely on their powerful language comprehension abilities to follow human-provided instructions and provide corresponding responses. Typically, LLMs can only respond within the text input provided by the user, which is insufficient because human interaction with the world involves multiple channels, such as visual and textual. To this end, recent

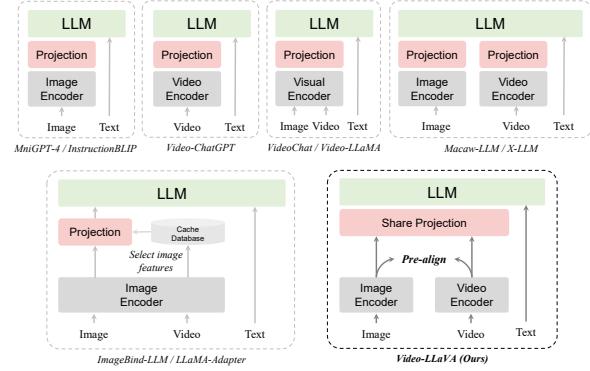


Figure 1: **Comparing Different LVLM Paradigms.** Video-LLaVA aligns images and videos before projection, allowing LLM to learn from a unified visual representation and endowing LLM with the ability to comprehend both images and videos simultaneously.

works (Ye et al., 2023; Zhu et al., 2023b; Alayrac et al., 2022) have mapped images into text-like tokens, enabling LLMs to emerge with the ability to comprehend images. Despite their effectiveness, empowering LLMs to understand videos is more challenging than image-only comprehension tasks. Nevertheless, recent work (Maaz et al., 2023; Li et al., 2023c; Zhang et al., 2023a) has made initial strides in enabling interactions between video and language.

However, most current LVLMs (Li et al., 2023b; Dai et al., 2023; Luo et al., 2023; Li et al., 2023a; Yin et al., 2023; Fu et al., 2023) can primarily handle a single visual modality, either image-language or video-language. We compare different LVLM paradigms as shown in Figure 1, where VideoChat (Li et al., 2023c) and Video-LLaMA (Zhang et al., 2023a) utilize a share visual encoder to handle both images and videos. However, due to the inherent differences in the media types of images and videos, it is challenging to learn a unified representation, and the performance falls significantly behind that of the specialized video expert model, Video-ChatGPT. Therefore,

X-LLM (Chen et al., 2023) and Macaw-LLM (Lyu et al., 2023) allocate a modality-specific encoder for each modality, attempting to enable a LLM to comprehend images or videos through several projection layers. But their performances are inferior to dedicated video expert models such as Video-ChatGPT (Maaz et al., 2023). We attribute this phenomenon to the lack of *alignment before projection*. Because image features and video features reside in their own spaces, this poses a challenge for a LLM to learn their interactions from several poor projection layers. Some similar phenomenon such as *alignment before fusion* has been discussed by ALBEF (Li et al., 2021) and ViLT (Kim et al., 2021) in multi-modal models. More recently, ImageBind-LLM (Han et al., 2023) focuses on enabling the LLM to simultaneously process multiple modal inputs by pre-aligning each modality to a common feature space (Girdhar et al., 2023). Based on a large image-language model, ImageBind-LLM converts other modalities into the most similar image features by retrieving from a training-free image cached database. However, the indirect alignment approach of ImageBind-LLM may lead to performance degradation, and the LLM has no knowledge of actual video data.

In this work, we introduce **Video-LLaVA**, a simple but powerful baseline for the LVLM simultaneously handling both images and videos. Specifically, As shown in Figure 1, Video-LLaVA initially aligns the representations of images and videos to a unified visual feature space. Since the visual representations are already aligned prior to projection, we employ a shared projection layer to map the unified visual representation for the LLM. To enhance computational efficiency, Video-LLaVA undergoes joint training of images and videos, achieving remarkable results with 1 training epoch.

As a result, The proposed Video-LLaVA greatly enhances the ability of the LLM to simultaneously understand both images and videos. For image understanding, Video-LLaVA surpasses advanced LVLMs such as mPLUG-owl-7B and InstructBLIP-7B in 5 image benchmarks. Additionally, utilizing 4 benchmark toolkits for a more comprehensive evaluation, Video-LLaVA-7B even outperforms IDEFICS-80B by 6.4% in MMBench. Moreover, similar trends can be observed in video understanding, where Video-LLaVA surpasses Video-ChatGPT by 5.8%, 9.9%, 18.6%, and 10.1% respectively on the MSVD, MSRVTT, TGIF, and ActivityNet video question-answering datasets. Ex-

tensive ablation experiments demonstrate that alignment before projection yields greater benefits. Additionally, joint training of images and videos can facilitate a unified visual representation in LLM comprehension.

We summarize our primary contributions as follows:

- We introduce **Video-LLaVA**, a powerful LVLM baseline. During the training process, Video-LLaVA binds visual signals to the language feature space, unifying visual representations, and proposes a solution to align before projection. We enable an LLM to perform visual reasoning capabilities on both images and videos simultaneously.
- Extensive experiments demonstrate that a unified visual representation benefits LLMs in learning to simultaneously handle both images and videos, validating the complementarity of modalities, showcasing significant superiority when compared to models specifically designed for either images or videos.

2 Related Work

2.1 Large Language Models

When the well-known commercial model ChatGPT (OpenAI, 2023) was introduced, the The AI community released open-source Large Language Models (LLMs) by instruction tuning and increasing model sizes. These include LLaMA (Touvron et al., 2023a), Vicuna (Chiang et al., 2023), Alpaca (Taori et al., 2023), and more recently, LLaMA 2 (Touvron et al., 2023b). These models are tuned with instruction sets to emulate conversations between humans and AI assistants. Furthermore, InstructGPT (Ouyang et al., 2022) is trained based on GPT-3 (Brown et al., 2020) with 175 billion parameters through aligning with human preferences. However, LLMs can only interact within text. In this work, we introduce Video-LLaVA, which builds upon the powerful reasoning capabilities of LLM to extend modality interactions to images and videos.

2.2 Large Vision-Language Models

When extending LLMs to multi-modal, especially involving images and videos, the main approaches can be categorized into two types in Table 1: *i*) treating LLM as a scheduler, *ii*) treating LLM as a decoder.

Table 1: **Comparison between different Large Vision-Language Models.** For methods that treat LLMs as scheduler, they do not require pre-alignment and joint training.

Methods	Image	Video	Pre-aligned	Joint training
<i>LLMs as scheduler</i>				
VisualChatGPT (Wu et al., 2023)	✓	✗	-	-
HuggingGPT (Shen et al., 2023)	✓	✗	-	-
MM-REACT (Yang et al., 2023)	✓	✓	-	-
ViperGPT (Surís et al., 2023)	✓	✓	-	-
<i>LLMs as decoder</i>				
Mini-GPT4 (Zhu et al., 2023b)	✓	✗	-	✗
LLaVA (Liu et al., 2023b)	✓	✗	-	✗
Video-ChatGPT (Maaz et al., 2023)	✗	✓	-	✗
VideoChat (Li et al., 2023c)	✓	✓	✗	✓
Video-LLaMA (Zhang et al., 2023a)	✓	✓	✗	✓
ImageBind-LLM (Han et al., 2023)	✓	✓	✓	✗
Video-LLaVA (Ours)	✓	✓	✓	✓

2.2.1 LLMs as scheduler

In the scheduler-based methods, various visual models are treated as plug-and-play modules. LLM schedules them according to the specific visual task requirements, like the assembly of building blocks. Some of these methods focus on images, such as VisualChatGPT (Wu et al., 2023) and HuggingGPT (Shen et al., 2023), while MM-REACT (Yang et al., 2023) and ViperGPT (Surís et al., 2023) can also handle videos. A key characteristic of these scheduler-based LVLMs is that they do not require end-to-end training, hence eliminating the need for pre-alignment and joint training of each modality.

2.2.2 LLMs as decoder

Regarding the approach of treating LLM as a decoder, this is our primary focus. MiniGPT-4 (Zhu et al., 2023b) aligns image tokens to the input of the large language model through several linear projection layers. However, this alignment is weak and lacks feedback from human instructions. Subsequently, mPLUG-Owl (Ye et al., 2023) adopts a two-stage training approach. In the first stage, images are aligned with language using an autoregressive pretraining style, and the second stage involves instruction tuning through using a human instruction dataset. With the increasing scale of large language model backends, approaches such as InstructBLIP (Dai et al., 2023) and LLaVA series (Liu et al., 2023b,a; Lin et al., 2024) collect the larger human instruction datasets to train a larger LVLMs (13B parameters). Each answer of

instruction datasets strictly follow to the given instructions. Then they undergo end-to-end training using human instruction datasets, enabling the LLM with visual reasoning capabilities. Moreover, Video-ChatGPT (Maaz et al., 2023) design a 100k video instruction dataset, successfully empowering LLMs to comprehend videos. VideoChat (Li et al., 2023c) and Video-LLaMA (Zhang et al., 2023a) achieve this by conducting joint training, allowing LLMs to simultaneously handle images and videos. Expanding LLMs to additional visual modalities typically requires pre-alignment, as seen in LLaMA-Adapter (Zhang et al., 2023b; Gao et al., 2023) and ImageBind-LLM (Han et al., 2023). They bind other modalities to the image space through ImageBind’s (Girdhar et al., 2023) modality encoder. These models have demonstrated that a unified feature space is advantageous for enhancing LLM’s multi-modal reasoning capabilities. Distinguished from prior work, Video-LLaVA not only pre-aligns image and video features but also conducts joint training of images and videos, facilitating LLMs in learning multi-modal reasoning capabilities from a unified visual representation.

3 Video-LLaVA

3.1 Model Structure

3.1.1 Framework Overview

As shown in Figure 2, Video-LLaVA consists of LanguageBind encoders f_V (Zhu et al., 2023a) to extract features from the raw visual signal (im-

ages or videos), a large language model f_L such as Vicuna, visual projection layers f_P and a word embedding layer f_T . We initially obtain visual features using LanguageBind encoders. LanguageBind encoders are capable of mapping different modalities into the textual feature space, thereby providing us with a unified visual representation. Subsequently, the unified visual representation is encoded by shared projection layers, which is then combined with tokenized textual queries and fed into a large language model to generate corresponding responses.

3.1.2 United Visual Representation

Our goal is to map images and videos into a shared feature space to enable the large language model to learn from a unified visual representation. We assume that the same information can be conveyed through multiple media. For example, a running dog can be expressed through language, a image or a video simultaneously. Therefore, we can compress information from different modalities into a common feature space, allowing the model to extract information from a dense feature space, facilitating modality interactions and complementarity. Hence, we chose the modality encoders from LanguageBind (Zhu et al., 2023a), which align images and videos with the textual feature space.

3.1.3 Alignment Before Projection

Specifically, LanguageBind initializes from OpenCLIP (Ilharco et al., 2021), naturally aligning images and language in a shared feature space. Subsequently, it aligns video representations to the language space using 3 million video-text pairs from VIDAL-10M (Zhu et al., 2023a). By sharing a language feature space, the image and video representations ultimately converge into a unified visual feature space, which we refer to as emergent alignment of images and videos. Therefore, our video encoder and image encoder are initialized from the LanguageBind encoders zoo, pre-aligning the inputs for LLM and reducing the gap between representations of different visual signals. The unified visual representation is fed into LLM after passing through a shared projection layer.

3.2 Training Pipeline

Overall, the process of generating responses by Video-LLaVA is similar to that of a large language model (GPT series). Given a textual input \mathbf{X}_T and visual signals \mathbf{X}_V , the input signals are encoded

into a sequence of tokens according to Equation 1. By maximizing the likelihood probability in Equation 2, the model ultimately achieves multi-modal understanding capabilities.

$$\mathbf{Z}_T = f_T(\mathbf{X}_T), \mathbf{Z}_V = f_P(f_V(\mathbf{X}_V)) \quad (1)$$

$$p(\mathbf{X}_A | \mathbf{X}_V, \mathbf{X}_T) = \prod_{i=1}^L p_\theta(\mathbf{X}_A^{[i]} | \mathbf{Z}_V, \mathbf{Z}_T^{[1:i-1]}) \quad (2)$$

where L is the length of the generated sequence \mathbf{X}_A , and θ is a trainable parameter. We dynamically conduct joint training on images and videos, wherein a single batch contains both image and video samples simultaneously.

3.2.1 Understanding Training

At this stage, the model is required to acquire the ability to interpret visual signals within an extensive image/video-text pair dataset. Each visual signal corresponds to a single round of conversation data $(\mathbf{X}_q, \mathbf{X}_a)$, where $\mathbf{X}_T = \mathbf{X}_q$ and \mathbf{X}_a is the ground truth. The training objective of this stage is the original auto-regressive loss, where the model learns the basic ability to view the vision. We freeze the other parameters of the model during this process.

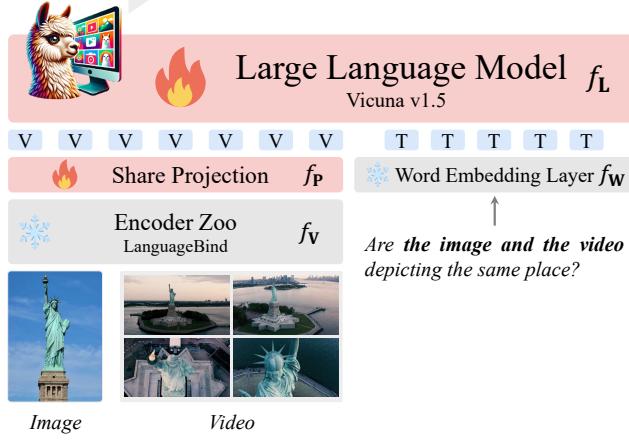
3.2.2 Instruction Tuning

In this stage, the model is required to provide responses corresponding to different instructions. These instructions often involve more complex visual comprehension tasks, rather than just describing visual signals. Note that the conversation data $(\mathbf{X}_q^1, \mathbf{X}_a^1, \dots, \mathbf{X}_q^N, \mathbf{X}_a^N)$ consists of multiple rounds.

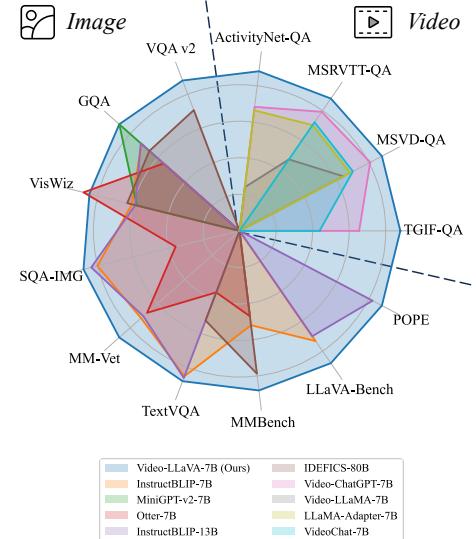
$$\mathbf{X}_T^r = \begin{cases} \mathbf{X}_q^1, & r = 1 \\ \text{Concat}(\mathbf{X}_q^{r-1}, \mathbf{X}_a^{r-1}, \mathbf{X}_q^r), & r > 1 \end{cases} \quad (3)$$

where r represents the round number. As shown in Equation 3, when $r > 1$ we concatenate the conversations from all previous rounds with the current instruction as the input for this round. The training objective remains the same as in the previous stage. After this stage, the model learns to generate corresponding responses based on different instructions and requests. The LLM are also involved in training at this stage.

Yes, the image and the video are depicting the same place. **The video shows the statue of liberty from different angles**, while **the image shows a close-up of the statue**. Both the video and the image capture the beauty and grandeur of the statue of liberty.



(a) Illustration of Video-LLaVA



(b) Performance comparison

Figure 2: **Training framework and performance.** Video-LLaVA exhibits remarkable interactive capabilities between images and videos, despite the absence of image-video pairs in the dataset. (a) The Video-LLaVA framework demonstrates a data flow that generates corresponding responses based on input instructions. (b) Video-LLaVA achieves superior performances on a broad range of 15 datasets across image and video.

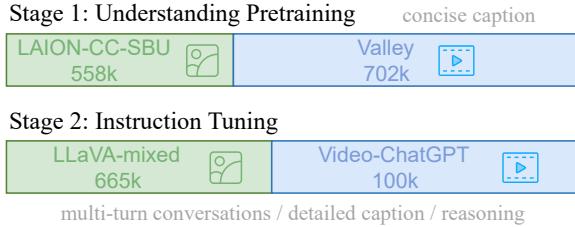


Figure 3: **Data composition for training Video-LLaVA.** The dataset for stage 1 consists of single-turn conversation, focusing on concise visual descriptions. In stage 2, the dataset comprises multi-turn conversations, emphasizing complex visual reasoning abilities.

4 Experiments

4.1 Experimental Setup

4.1.1 Data Details

In 3, for the first stage of understanding pretraining, we use a subset of 558K LAION-CC-SBU image-text pairs with BLIP (Li et al., 2022) captions, which is sourced from CC3M (Sharma et al., 2018) and filtered by LLaVA (Liu et al., 2023b). The video-text pairs are derived from a subset provided by Valley (Luo et al., 2023), and we have access to 702k out of a total of 703k pairs, originating from WebVid (Bain et al., 2021). For the stage of instruction tuning, We gathered instructional datasets from two sources, including a 665k image-text instruction dataset from LLaVA 1.5 (Liu et al., 2023a) and a 100k video-text instruction dataset from Video-ChatGPT (Maaz et al., 2023).

4.1.2 Model Settings

We employ Vicuna-7B v1.5 as the large language model. The visual encoders are derived from LanguageBind, initialized from OpenCLIP-L/14. The text tokenizer is sourced from LLaMA, with approximately 32,000 classes. The share projection layers consist of 2 fully connected layers with a GeLU (Hendrycks and Gimpel, 2016) activated function.

4.1.3 Training Details

In the training process, we resize and crop each image, resulting in a size of 224×224 for each processed image. We uniformly sample 8 frames from each video, and each frame undergoes image preprocessing. The data in each batch is a random combination of images and videos. In the first stage, we train for one epoch with a batch size of 256, using the AdamW optimizer with a cosine learning rate schedule. In the second stage, we reduce the batch size to 128. The initial learning rate for both stages is set to 1e-3, with a warmup ratio of 0.03. Additional hyper-parameter settings can be found in the appendix.

4.2 Quantitative Evaluation

4.2.1 Zero-shot Video Understanding

As shown in Table 2, we conduct a quantitative assessment of the video question-answering ca-

Table 2: **Comparison between different LVLMs on video reasoning benchmarks.** We employ ChatGPT-Assistant to evaluate the performance following Video-ChatGPT (Maaz et al., 2023). The version of ChatGPT is “gpt-3.5-turbo”.

Methods	LLM size	MSVD-QA		MSRVTT-QA		TGIF-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	1B	32.2	-	16.8	-	41.0	-	24.7	-
VideoChat	7B	56.3	2.8	45.0	2.5	34.4	2.3	-	2.2
LLaMA-Adapter	7B	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video-LLaMA	7B	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	7B	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
Chat-UniVi	7B	<u>65.0</u>	<u>3.6</u>	<u>54.6</u>	<u>3.1</u>	<u>60.3</u>	<u>3.4</u>	45.8	<u>3.2</u>
Video-LLaVA	7B	70.7	3.9	59.2	3.5	70.0	4.0	45.3	3.3

Table 3: **Comparison between different LVLMs on image understanding benchmarks.** “Res.”, “L”, “V” respectively represent the input image resolution, LLaMA (Touvron et al., 2023a) and Vicuna (Chiang et al., 2023). Benchmark names are abbreviated due to page limitations. VQA-v2 (Goyal et al., 2017); GQA (Hudson and Manning, 2019); VisWiz (Gurari et al., 2018); SQA^I: ScienceQA-IMG (Lu et al., 2022); VQA^T: TextVQA (Singh et al., 2019); POPE (Li et al., 2023d); MMB: MMBench (Liu et al., 2023c); LLaVA^W: LLaVA-Bench (In-the-Wild) (Liu et al., 2023b); MM-Vet (Yu et al., 2023). [†] donates that we reproduce LLaVA-1.5 with LanguageBind-Image encoder to compare fairly. ^{*} donates that there is some overlap in the training data.

Methods	LLM	Res.	Image Question Answering					Benchmark Toolkit			
			VQA ^{v2}	GQA	VisWiz	SQA ^I	VQA ^T	POPE	MMB	LLaVA ^W	MM-Vet
LLaVA-1.5	V-7B	336	-	62.0*	-	-	-	-	-	-	30.5
BLIP-2	V-13B	224	41.0	41.0	19.6	61.0	42.5	85.3	-	38.1	22.4
InstructBLIP	V-13B	224	-	49.5	33.4	63.1	50.7	78.9	-	58.2	25.6
IDEFICS-80B	L-65B	224	60.0	45.2	36.0	-	30.9	-	54.5	-	-
MiniGPT-4	L-7B	224	-	30.8	47.5	25.4	19.4	-	23.0	-	22.1
IDEFICS-9B	L-7B	224	50.9	38.4	35.5	-	25.9	-	48.2	-	-
mPLUG-Owl	L-7B	224	-	14.0	39.0	2.8	38.8	-	46.6	-	-
Otter	L-7B	224	-	38.1	50.0	27.2	21.2	-	32.6	-	24.6
InstructBLIP	V-7B	224	-	49.2	34.5	60.5	<u>50.1</u>	-	36.0	60.9	<u>26.2</u>
LLaVA-1.5 [†]	V-7B	224	72.3*	56.9*	47.8	67.9	49.2	83.3	<u>59.5</u>	<u>63.3</u>	25.7
Video-LLaVA	V-7B	224	74.7*	60.3*	48.1	66.4	51.8	84.4	60.9	73.1	32.0

pabilities of large video-language models on four datasets, including MSVD-QA (Chen and Dolan, 2011), MSRVTT-QA (Xu et al., 2016), TGIF-QA (Jang et al., 2017) and ActivityNet-QA (Yu et al., 2019). The evaluation pipeline for video understanding follows Video-ChatGPT. We report the accuracy and score, which is assessed using GPT-Assistant. Video-LLaVA consistently outperforms Video-ChatGPT in terms of question-answering accuracy, which is an advanced large video-language model. Moreover, Video-LLaVA surpasses the powerful baseline of Video-ChatGPT by 5.8%, 9.9%, 18.6%, and 10.1% on MSRVTT, MSVD, TGIF, and ActivityNet, respectively. Additionally, we conduct comparisons with the recent

SOTA model, Chat-UniVi (Jin et al., 2023). Despite Chat-UniVi utilizing more datasets such as MIMIC-IT (Li et al., 2023a), Video-LLaVA still demonstrate competitive results, surpassing Chat-UniVi on MSVD, MSRVTT, and TGIF datasets. In summary, these results validate Video-LLaVA’s ability to comprehend videos and provide contextually appropriate responses based on instructions.

4.2.2 Zero-shot Image Question-answering

As shown in Table 3, we evaluate our approach for image understanding on five academic image question-answering benchmarks. Compared to the state-of-the-art model InstructBLIP-7B, Video-LLaVA demonstrates powerful image understand-

Table 4: **Zero-shot object hallucination evaluation results** are reported for three POPE evaluation settings. “Yes” indicates the proportion of positive responses to the given question. \dagger donates that we reproduce LLaVA-1.5 with LanguageBind-Image encoder to compare fairly.

Methods	LLM	Adersarial			Popular			Random		
		Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes	Accuracy	F1-Score	Yes
MiniGPT-4	V-13B	66.6	71.4	66.7	68.3	72.2	64.1	77.8	78.9	54.8
InstructBLIP	V-13B	74.4	78.5	69.0	81.4	83.5	62.6	88.7	89.3	55.2
MM-GPT	L-7B	50.0	66.7	100.0	50.0	66.7	100.0	50.0	66.7	100.0
mPLUG-Owl	L-7B	50.7	66.8	98.7	50.9	66.9	98.6	54.0	66.4	95.6
Chat-UniVi	V-7B	55.6	68.7	91.6	56.4	69.0	90.8	73.9	79.3	74.6
LLaVA-1.5 \dagger	L-7B	84.3	83.2	43.5	<u>79.8</u>	<u>79.4</u>	48.0	<u>85.7</u>	<u>84.8</u>	43.0
Video-LLaVA	V-7B	<u>81.6</u>	<u>80.8</u>	45.8	85.3	84.0	42.1	86.2	85.2	42.0

ing capabilities, outperforming across all five question-answering benchmarks. Additionally, Video-LLaVA exhibits competitive results compared to several more powerful LVLMs, which are tuned based on 13B or 65B LLM, such as surpassing InstructBLIP-13B by 14.7% on VisWiz, highlighting its strong understanding ability in natural visual environments. Furthermore, to ensure a fair comparison, we replace the image encoder in LLaVA-1.5 with the LanguageBind-Image encoder, called LLaVA-1.5 \dagger . This demonstrates that the performance improvement observed in Video-LLaVA is not solely attributed to a stronger image encoder. Additional details can be found in Section 4.3.6.

Evaluation under Image Benchmark Toolkits
 Additionally, we evaluate LVLMs using several benchmark toolkits for visual instruction tuning. These benchmark toolkits provide a detailed assessment of the model’s capabilities through robust evaluation metrics. Video-LLaVA outperform InstructBLIP-7B by 24.9%, 12.2%, and 5.8% on MMBench, LLaVA-Bench, and MM-Vet, respectively. It is worth noting that Video-LLaVA-7B still demonstrates advanced performance compared to larger LLM models, surpassing InstructBLIP-13B by 6.4% on MM-Vet and IDEFICS-80B (Lau-rençon et al., 2023) by 6.4% on MMBench. These results demonstrate that Video-LLaVA exhibits a strong understanding of semantic aspects of scenes, enabling it to answer open-ended and free-form natural language questions about images.

4.2.3 Object Hallucination Evaluation

As shown in Table 4, we report evaluation results for zero-shot object hallucinations, utilizing a evaluation pipeline derived from a polling-based query method (Li et al., 2023d). Video-LLaVA demon-

strates competitive performance across three subsets: random, popular, and adversarial. Specifically, when compared to the 7B foundation model, Video-LLaVA consistently outperforms MM-GPT (Gong et al., 2023) across all three POPE hallucination evaluation subsets. Furthermore, when benchmarked against the larger 13B LLM, Video-LLaVA even surpasses Mini-GPT4 comprehensively. The successful performance of Video-LLaVA in object hallucination detection validates the consistency between unified visual representations and the generation of textual descriptions.

4.3 Ablation Results

4.3.1 Alignment Before Projection

To validate the performance degradation caused by separated visual representation, we conduct experiments to explore the performance of the LLM learning from different visual representations. We define the use of LanguageBind image encoder as unified visual representation while the MAE encoder (He et al., 2022) use separated visual representation, which is a well-known and effective image feature extractor. Additionally, since MAE do not interact with multi-modal inputs during the training process, we utilize CLIP-L/14, a model of the same size. While CLIP-L/14 exhibits strong multimodal understanding capabilities, it is not pre-aligned with the video encoder. Consequently, this results in a lack of uniformity in the visual features provided to LLM. We only replace the image encoder of the same scale and keep the LanguageBind video encoder.

4.3.2 For Video Understanding

Due to replacing the image encoder with the MAE encoder, the video features and image features are

Table 5: **Effect of alignment before projection on image.** “United” refers to the unified visual representation, while “Separated” refers to the separated visual representation. Benchmark names are abbreviated due to page limitations.

Methods	Image Question Answering					Benchmark Toolkit			
	VQA ^{v2}	GQA	VisWiz	SQA ^I	VQA ^T	POPE	MMB	LLaVA ^W	MM-Vet
Separated-MAE	66.0	55.4	42.5	65.0	44.2	80.8	45.7	35.9	20.0
Separated-CLIP	74.6	59.9	47.8	67.3	51.5	84.4	60.2	68.9	30.6
United	74.7	60.3	48.1	66.4	51.8	84.4	60.9	73.1	32.0
Δ Acc.	+0.1	+0.4	+0.3	-0.9	+0.3	+0.0	+0.7	+4.2	+1.4

Table 6: **Effect of joint training on video.** We evaluate on four video question-answering datasets. * denotes that we utilized only video data in both the first and second stages.

Methods	MSVD-QA		MSRVT-QA		TGIF-QA		ActivityNet-QA	
	Accuracy Score		Accuracy Score		Accuracy Score		Accuracy	Score
Video-LLaVA*	64.8	3.2	58.3	3.4	67.8	3.4	40.7	2.0
Joint with Image	70.7	3.9	59.2	3.5	70.0	4.0	45.3	3.3
Δ Acc.	+5.9	+0.7	+0.9	+0.1	+2.2	+0.6	+4.6	+1.3

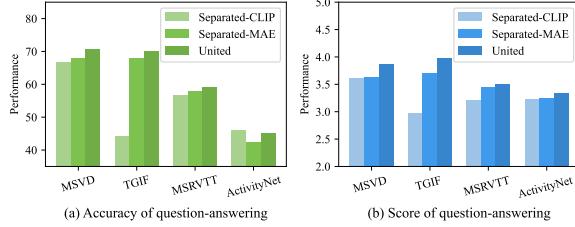


Figure 4: **Effect of alignment before projection on video.** We validate and report the accuracy and score on four video question-answering datasets.

no longer unified during LLM’s initial learning of visual representations. In Figure 4, compared to separated visual representation, the united visual representation significantly improves performance across 4 video question-answering datasets. Separated visual representations not only exhibit lower accuracy in question-answering, but also demonstrate a similar trend in answer scores. These results demonstrate that the unified visual representation can help the LLM further learn and understand videos.

4.3.3 For Image Understanding

The unified visual representation demonstrates strong performance, surpassing the separated visual representation comprehensively across 5 image question-answering datasets and 4 benchmark toolkits in Table 5. Additionally, we observe a significant margin of performance improvement in the unified visual representation on the MM-

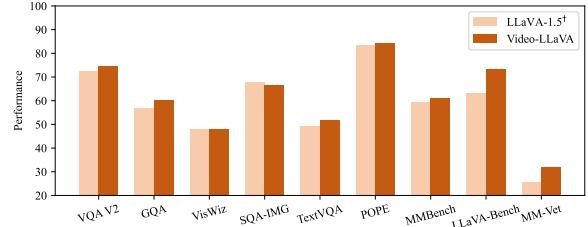


Figure 5: **Effect of joint training on image.** † denotes that we reproduce the results of LLaVA-1.5 at a resolution of 224×224 with LanguageBind-Image encoder for a fair comparison.

Bench, LLaVA-Bench, and MM-Vet benchmark toolkits. This highlights that the unified visual representation not only enhances performance in image question-answering but also provides benefits in other aspects of image understanding, such as reducing object hallucination and improving OCR capabilities.

4.3.4 Joint Training

This subsection aims to validate the complementarity of images and videos during joint training, which can mutually enhance the LLM’s understanding of images and videos based on a unified visual representation.

4.3.5 For Video Understanding

For comparing performance on video benchmarks, we remove image data during the training of Video-LLaVA, which is called Video-LLaVA*. We com-

pare with Video-LLaVA* to assess the performance gains from joint image training on video benchmarks. In Table 6, we evaluate our model on four video question-answering datasets. Compared to Video-LLaVA* without image in training, the model trained with joint images and videos achieves comprehensive improvements across all four video datasets. These results demonstrate that joint training of images and videos facilitates LLM’s understanding of visual representations.

4.3.6 For Image Understanding

When comparing performance on image benchmarks, it is challenging to find a image-based LVLM with the same configuration as Video-LLaVA. To address this, we replace the image encoder in LLaVA-1.5 with the LanguageBind-Image encoder and reproduce the results at a resolution of 224×224 by using the same training configuration, called LLaVA-1.5[†]. As shown in Figure 5, Compared to LLaVA-1.5[†], which utilizes the same image encoder configuration, we observe performance improvements in 8 out of 9 benchmarks, demonstrating mutual improvement in visual understanding. Video-LLaVA outperform LLaVA-1.5[†] in POPE, indicating that joint training with videos alleviates the object hallucination in images. The similar trend is observed on some other benchmark toolkits, such as LLaVA-Bench and MMBench, where video data significantly improves LLM’s performance in complex reasoning and image conversation tasks.

5 Limitation and Future Directions

5.1 Limitation

While Video-LLaVA exhibits strong competitiveness in both images and videos, we still observed some limitations of Video-LLaVA. To begin with, Video-LLaVA performs moderately in understanding long videos. In Table 2, Chat-UniVi surpasses 0.5 on ActivityNet-QA because Video-LLaVA only utilizes uniformly sampled 8 frames to comprehend the video, which results in the loss of detailed information from long videos. Additionally, training Video-LLaVA is computationally expensive, requiring 3-4 days to complete the training process on 8 A100-80G GPUs.

5.2 Future Directions

In the future, We maybe can explore more efficient shared projection mode that can compress tokens

while preserving data features. This would support Video-LLaVA in better understanding long videos. Besides, Video-LLaVA can serve as a baseline to extend to additional visual-related modalities, such as depth and infrared images. Additionally, we could explore how to incorporate timestamp embeddings effectively, enabling large visual-language models to answer questions related to temporal relationships.

6 Conclusion

In this work, we introduce Video-LLaVA, a simple but powerful large visual-language baseline model. We propose a novel framework to address the issue of misalignment before projection, utilizing a LanguageBind encoder to pre-bind visual signals into the language feature space. To enable a LLM to comprehend both images and videos simultaneously, we conduct joint training on images and videos, allowing the LLM to learn multi-modal interactions from a unified visual representation. Extensive experiments demonstrate that joint training on images and videos mutually benefits performance. Furthermore, we validate that aligning visual representations before projection aids LLM learning. Remarkably, LLM, after learning from a unified visual representation, exhibits the remarkable ability to simultaneously engage with both images and videos, showcasing a powerful comprehension of unified visual concepts. These results collectively demonstrate the effectiveness of the Video-LLaVA training framework. As a unified visual training framework, the performance of Video-LLaVA even surpasses that of expert models designed specifically for images or videos.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China (No. 62202014, 62332002, 62425101), Shenzhen Basic Research Program (No.JCYJ20220813151736001).

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Max Bain, Arsha Nagrani, Gü̈l Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.
- Bin Bi, Chenliang Li, Chen Wu, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2020. Palm: Pre-training an autoencoding&autoregressive language model for context-conditioned generation. *arXiv preprint arXiv:2004.07159*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.
- Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023).
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint, arXiv:2305.06500*.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.
- Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.
- Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.
- Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

- Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. 2023. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint, arXiv:2306.16527*.
- Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.
- Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.
- OpenAI. 2023. Gpt-4 technical report. *Preprint, arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. 2023. Hugging-gpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

- Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. Mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueteng Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.
- Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhai Jiang, Junwu Zhang, Zongwei Li, et al. 2023a. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Example Appendix

A.1 Training Setting

We show some training settings as shown in Table 7. video encoder and image encoder are not trained in both stages. The projection layer consists of 2 linear layers with a GeLU (Hendrycks and Gimpel, 2016) activation function between them. Image and video share the projection layer.

Table 7: Training setting.

Config	Pretraining	Instruction tuning
Video encoder	LanguageBind-Video-LoRA	
Image encoder	LanguageBind-Image	
Optimizer	AdamW	
Deepspeed	Zero2	
Epochs	1	
Vision select layer	-2	
Weight decay	0.0	
Warmup ratio	0.03	
Learning rate schedule	cosine decay	
Learning rate	1e-3	2e-5
Batch size	256	128

A.2 Exhibition Board

We show some **unselected** samples here, and these videos are sourced from Video-ChatGPT (Maaz et al., 2023).



[Input Instruction] Describe the video.

The video shows a young girl performing a gymnastics routine on a balance beam. She executes several flips and tricks before dismounting and landing on a mat.

[Input Instruction] Is there a flag in the background?

Yes, there is a flag in the background of the video.



[Input Instruction] What is the video about?

The video is about a man who is seen spinning and throwing a discus in a field. He then walks away while people measure his throw.



[Input Instruction] What activities are the woman and the dog engaged in the video?

The woman and the dog are playing frisbee in the yard. The woman throws the frisbee, and the dog chases after it.



[Input Instruction] What is the video about?

The video shows two men arm wrestling on a table while being watched by a group of people.



[Input Instruction] Step-by-step describe the cooking recipe shown in the video.

The video shows a person cooking a dish of noodles. They start by boiling water in a pot and adding noodles to it. Then, they add oil and some ingredients to the pot. After that, they mix the noodles and the ingredients together. Finally, they serve the dish on a plate.

Figure 6: Samples of Video-LLaVA in video understanding.



[Input Instruction] Why is this video funny?

The video is funny because the baby is playing with a Wii remote, which is not a toy typically associated with babies. The baby's actions are also humorous, such as pointing to the camera and making faces, which adds to the overall amusement of the video.



[Input Instruction] Can you identify any safety hazards in this video?

Yes, the baby is playing with a toy truck and a toy car, which could pose a safety risk if they are not supervised. However, the baby is sitting on the floor, so there is less risk of injury compared to if they were playing with the toys on a higher surface.



[Input Instruction] Create a short fairy tale with a moral lesson inspired by the video.

Once upon a time, in a kingdom far away, there lived a beautiful princess named Rose. She was known for her kindness and beauty, and everyone loved her. One day, a handsome prince named Leo came to visit the kingdom. He was fascinated by Rose's beauty and decided to stay in the kingdom. However, Rose's jealous maid, Lily, did not like this and Leo. Lily gave her sister, the evil queen, a potion that made her fall in love with Leo. The kingdom celebrated their love and they lived happily ever after. The moral of the story is that true love conquers all and that jealousy can lead to harmful consequences.



[Input Instruction] Describe the video.

The video shows a person preparing a dish with peppers and oranges. They slice the peppers and oranges and lay them out on a table. The person then proceeds to cut the peppers and oranges into strips and lay them out on a piece of wax paper. They use a knife to cut the peppers and oranges into strips and lay them out on a piece of wax paper. The person then uses a spoon to mix the peppers and oranges together. The video ends with the person cutting the peppers and oranges into strips and laying them out on a piece of wax paper.



[Input Instruction] What is the main focus of the video?

The main focus of the video is the process of creating a painting of a tree using various techniques and colors. The video shows the artist's hand holding a paintbrush and using it to create the tree, as well as the use of different colors and techniques to create the painting.



[Input Instruction] What does the woman use to split the logs and how does she do it?

The woman uses an ax to split the logs. She holds the ax in her hand and swings it down onto the log. She repeats this process several times, splitting the log into smaller pieces.

Figure 7: Samples of Video-LLaVA in video understanding.



Figure 8: Samples of Video-LLaVA in video understanding.

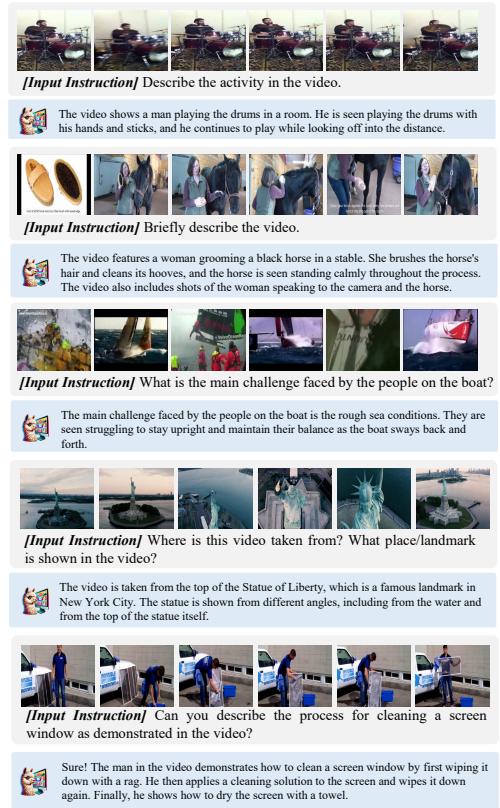


Figure 9: Samples of Video-LLaVA in video understanding.