PAVE: Patching and Adapting Video Large Language Models

Zhuoming Liu[1], Yiquan Li[1], Khoi Duc Nguyen[1], Yiwu Zhong[2], Yin Li[1]
[1]University of Wisconsin-Madison [2]The Chinese University of Hong Kong

In this supplement, we (1) show additional experiment results on small Video LLM and multiple-view video understanding (Section A); (2) describe additional implementation details (Section B); (3) include additional visualization of the question-answering results (Section C).

A. Additional Experiment Results

A.1. Results with small Video LLM

We now present additional experiment results of PAVE with LLaVA-OneVision 0.5B models for audio-visual QA and 3D QA. Table 1 and Table 2 show the results. PAVE consistently improves the 0.5B and 7B Video LLM's performance by a large margin across both settings. This indicates that PAVE effectively leverages additional information when adapting pre-trained Video LLMs into new settings.

A.2. Results on Enhanced Video Understanding

We present PAVE's result on additional benchmarks in the enhanced video understanding setting. Table 3 shows PAVE's results in the enhanced video understanding setting with additional benchmarks. PAVE demonstrates a substantial performance gain on VideoMME (w-subtitles). However, we observe only marginal or no improvement on ActivityNet-QA, EgoSchema, NextQA, and Perception-Test. We hypothesize that this discrepancy may be due to: (1) domain shift—our training data primarily consists of third-person view videos, which may lead to a performance drop in EgoSchema, and (2) the nature of the benchmark questions, which may not require densely temporal information for reasoning.

A.3. Results on Multi-view Video Understanding

Motivation and task set up. Understanding human activity from video is crucial in many real-world applications, such as augmented reality and robotic learning. Based on the perspective, videos can be broadly classified into ego-centric and exo-centric views. Ego-centric videos capture first-person interactions, focusing on close-up hand-object interactions, while exo-centric videos provide a third-person perspective, recording full-body postures and the surrounding environment. Both perspectives are essential for comprehensive human action understanding. Different from the audio-visual QA and 3D QA, where the side-channel information comes from other modalities, in this context, PAVE regards exo-centric videos as side-channel information and integrates it with ego-centric video to adapt the Video LLMs for multi-view video understanding.

Training data. We use the training set from the Ego-Exo4D demonstrator proficiency estimation benchmark [8] as our training data, which consists of 1,904 question-answer pairs. Each pair is associated with one ego-centric video and four exo-centric videos. The task requires the model to classify human action proficiency into one of four categories: Novice, Early Expert, Intermediate Expert, or Late Expert, based on both ego- and exo-centric videos. However, only 1,656 question-answer pairs include the corresponding videos, as the videos for the remaining pairs could not be downloaded due to privacy issues.

Implementation details. Considering the exo- and ego-centric videos are synchronized along the temporal axis, we sample 32 frames for each of the exo-centric videos. To keep the encoding procedure consistent between the ego- and exo-video, we use the same preprocessing of the LLaVA-OneVision to reshape and crop the video frames. We use SigLIP [20] as the visual encoder and it encodes and downsamples each frame into 196 tokens. We pre-extract the exo-video feature tokens offline to accelerate the training. We build PAVE on top of LLaVA-OneVision [9] and train the model for 2 epochs.

Evaluation benchmark. We use the validation set of the Ego-Exo4D [8] demonstrator proficiency estimation benchmark for evaluation and report accuracy as the metric. It contains 466 questions and each of the questions is paired with 1 ego-centric video and 4 exo-centric videos.

Baselines. We use the TimeSFormer (Ego+Exo) from Ego-Exo4D [8] as our baseline. We also include a baseline that directly fine-tunes the LLaVA-OneVision with LoRA on the training set without using the exo-centric videos, denoted as LLaVA-OV-7B-FT. This baseline allows us to assess whether PAVE can effectively utilize supplementary information.

Results. Table 4 shows the results of PAVE. Compared with the LLaVA-OV-7B-FT, PAVE-7B achieves about 14.4% im-

Method	AVSD [1] CIDEr	AVQA [17] Acc.	MUSIC-AVQA [10] Audio Acc.	Visual Acc.	Audio-Visual Acc.	Overall Acc.	TFLOPs	Total / Trainable Params
Zero-shot LMMs								
LLaVA-OV-0.5B [9]	65.1	77.4	60.0	57.1	48.5	52.8	8.01	0.9B / -
LLaVA-OV-7B [9]	70.6	85.6	68.8	70.6	52.8	60.4	98.53	8.2B / -
Task-specific models								
LLaVA-OV-0.5B-FT	117.6	86.4	69.6	76.3	62.8	67.6	8.01	0.9B / 35.2M
LLaVA-OV-7B-FT	124.9	90.8	75.4	89.3	72.3	77.4	98.53	8.2B / 161.5M
PAVE-0.5B (w/ audio)	134.5	90.4	77.3	89.8	74.1	78.8	8.08	0.9B / 41.4M
PAVE-7B (w/ audio)	152.9	93.8	79.7	93.0	78.0	82.3	98.63	8.2B / 170.5M

Table 1. Additional result of PAVE on the audio-visual understanding tasks with audio as additional information.

Method	ScanQA [2] C	B-4	M	R	EM@1	SQA3D[14] EM@1	TFLOPs	Total / Trainable Params
Zero-shot LMMs								
LLaVA-OV-0.5B [9]	17.2	1.2	13.7	18.4	0.2 (28.0)	0.8 (43.0)	8.01	0.9B /-
LLaVA-OV-7B [9]	91.0	5.3	18.2	45.9	26.7 (44.3)	8.3 (50.7)	98.53	8.2B / -
Task-specific models								
LLaVA-OV-0.5B-FT	70.5	6.5	14.3	36.9	20.5 (36.3)	44.1 (45.7)	8.01	0.9B / 35.2M
LLaVA-OV-7B-FT	95.1	13.5	19.1	47.4	27.4 (46.3)	55.8 (58.1)	98.53	8.2B / 161.5M
PAVE-0.5B (w/ 3D info)	84.2	13.1	17.0	42.1	23.1 (40.0)	51.1 (52.8)	8.13	0.9B / 41.4M
PAVE-7B (w/ 3D info)	103.4	16.0	19.9	49.0	29.1 (48.5)	59.0 (61.4)	98.68	8.2B / 170.5M

Table 2. Additional result of PAVE on the 3DQA tasks with 3D information as additional information.

provement by adding only 9M parameters and 0.17 TFLOPs during inference. This big improvement indicates that the exo-centric videos provide crucial additional information for human action understanding. Moreover, PAVE achieves state-of-the-art performance on the demonstrator proficiency estimation benchmark, substantiating that PAVE can adapt a pre-trained Video LLM to an unseen setting by leveraging supplementary information.

B. Implementation and Experiment Details

We first describe the general implementation detail of the PAVE in Section B.1. Then, we describe the experiment details for 3 settings considered in the main paper, including audio-visual QA (Section B.2), 3DQA (Section B.3), and enhancing video QA (Section B.4). We also demonstrate how we calculate the Flops for the model in Section B.5.

B.1. Implementation Detail of PAVE

Inside the temporal-aligned cross-attention layer, we add rotary position embedding to the query and key tokens. Specifically, we apply different rotary positional embedding according to the layout of side-channel tokens $\mathbf{z}^s$. We mainly consider two types of $\mathbf{z}^s$: (a) $\{\mathbf{z}^s\}$ includes both spatial and temporal dimensions, such as tokens from video backbones or from a 3D backbone; and (b) $\{\mathbf{z}^s\}$ only contains temporal dimension, such as audio tokens. For the first case, we will add 3D rotary positional embedding (along the temporal, height, and width dimensions). For the second case, we will only add rotary positional embedding along the temporal axis. After cross-attention, we use a two-layer MLP, followed by a layer norm. After the PAVE layers, we add another two-layer MLP, followed by a layer norm, as the adapter. We initialize the γ in the layer norm to zero.

B.2. Audio-Visual QA

In this setting, the input of the PAVE has two parts: (1) the visual tokens $\mathbf{z}^v$ from the Video LLM's visual encoder, and (2) the audio tokens $\mathbf{z}^s$ from a side-channel signal encoder.

Visual Encoder. For $\mathbf{z}^v$, we follow the default setting used in LLaVA-OneVision [9]. We uniformly sample 32 frames from the video and use the same preprocessing of the LLaVA-OneVision to reshape and crop the video frames. We use SigLIP [20] as the visual encoder and it encodes and downsamples each frame into 196 tokens.

Side-Channel Signal Encoder. For $\mathbf{z}^s$, we follow the preprocessing step of ImageBind [7], which resamples the audio at 16KHz. We segment the audio into overlapping 2-second clips with a 1-second stride and encode each clip using the audio encoder of ImageBind. This process generates a 1024-dimensional audio token for every 1 second of the audio signal. Since we do not fine-tune the audio encoder, we extract the audio feature tokens offline in order to accelerate the training.

Network Architecture. For the PAVE design, we use 2 cross-attention layers with hidden dimension 512 and have 4 attention heads. For LoRA layers in the LLM, we use LoRA_r = 64 and LoRA_α = 16.

Method	ActivtityNet-QA	EgoSchema	NextQA	PerceptionTest	VideoMME (w-subs)	FLOPs (TB)	Total / Trainable Params
LLaVA-OV-0.5B [9]	50.5	26.8	57.2	49.2	43.5	8.01	0.9B / -
LLaVA-OV-7B [9]	56.6	60.1	79.4	57.1	61.5	98.53	8.2B / -
PAVE-0.5B (w/ video feature)	50.6	27.1	56.1	48.8	48.6	8.08	0.9B / 41.4M
PAVE-7B (w/ video feature)	57.1	57.4	79.6	56.0	62.9	98.63	8.2B / 170.5M

Table 3. Result of PAVE on the additional benchmarks in enhanced video understanding setting. PAVE uses densely sampled video frames as additional information.

Model	Acc.	FLOPs (TB)	Total / Trainable Params
Zero-shot LMMs			
LLaVA-OV-0.5B	23.6	8.01	0.9B / -
LLaVA-OV-7B	23.6	98.53	-
Task-specific models			
LLaVA-OV-0.5B-FT	28.2	8.01	0.9B / 35.2M
LLaVA-OV-7B-FT	29.8	98.53	8.2B / 161.5M
TimeSFormer (Ego+Exo)* [8]	43.7	-	-
PAVE-0.5B	32.4	8.15	0.9B / 41.4M
PAVE-7B	44.2	98.70	8.2B / 170.5M

Table 4. Performance of PAVE on multi-view video understanding with Ego-Exo4D Demonstrator Proficiency benchmark. LLaVA-OV-7B-FT refers to directly fine-tuning the LLaVA-OneVision on the training set. Our model achieves state-of-the-art performance by only adding a small amount of parameters and FLOPs. * means this baseline may use more training data than PAVE because some of the videos are unavailable to us.

Training Details. For training, we use AdamW [13] optimizer with a linear warmup using the first 3% of iterations. We use the cosine annealing learning rate during the training. We set the base learning rate as 2e-5 and the batch size as 32. All the experiments are run on 2 A100 80G GPUs.

Training data. We choose the open-end QA dataset AVSD [1], and closed-end QA dataset AVQA [17] and Music-AVQA [10] as training dataset. AVSD contains 79k question-answer pairs across 7,985 videos with each paired 10 questions. AVQA has 40k question-answer pairs coupled with 40k Videos. Music-AVQA consists of 32k question-answer pairs and 9277 videos.

Evaluation benchmark. We follow the protocol in previous works [15, 18] to evaluate PAVE. For AVSD, we use the AVSD@DSTC7 test split and report CIDEr score as the metric. This benchmark consists of 1,000 audio-visual questions. We use COCO API [12] to calculate the CIDEr score between the model predictions and the ground truth answers. For AVQA, we evaluate PAVE on the eval split and report the accuracy as the metric. This benchmark contains 17k questions that require reasoning based on audio and visual information. For Music-AVQA, we evaluate PAVE on the test split and report the accuracy as the metric. This benchmark contains 9185 questions, which can be categorized into visual, audio, and audio-visual questions.

B.3. 3DQA

In this setting, the input of the PAVE consists of two parts: (1) the visual tokens $\mathbf{z}^v$ from the Video LLM's visual encoder, and (2) the 3D tokens $\mathbf{z}^s$ from a side-channel signal encoder.

Visual Encoder. For $\mathbf{z}^v$, we use the same setting as the one in Section B.2.

Side-Channel Signal Encoder. For encoding the side-channels information into $\mathbf{z}^s$, we utilize the 3D encoder which contains two parts 1. a visual encoder which encodes the RGB frames into visual feature tokens. 2. a spatial embedding that adds the encoded 3D information on the visual feature tokens. We uniformly extract 32 RGB-D frames from the scan and use ViT [4] to extract the visual features from the RGB frames. We then add spatial embeddings to visual features following the LLaVA-3D [24] by making use of the depth information and the camera pose. It generates 576 tokens for each frame, with a token dimension of 1024. We pre-extract the 3D feature to accelerate the training.

Network Architecture and Training Details. For the PAVE design and the training configuration, we use the same hyper-parameters used in Section B.2.

Training data. For 3D QA tasks, we consider ScanQA [2] and SQA3D [14]. ScanQA and SQA3D contain 25K and 26K training question-answer pairs, respectively. They share the same scanning data set which contains 562 3D scanning from ScanNet [3].

Evaluation benchmark. We report our model performance on the ScanQA validation set, which contains 4,675 questions covering both object position reasoning and object recognition, and the SQA3D test set with 3519 questions, which consists of 5 different types of questions. Following previous work [24], we report the CIDEr (C), BLEU-4 (B-4), METEOR (M), ROUGE(R), and top-1 Exact Match (EM@1) metrics on ScanQA and report EM@1 on SQA3D. We use the evaluation pipeline set up by LLaVA-3D to evaluate our model on ScanQA and SQA3D.

B.4. Enhancing Video QA

In this setting, the input of the PAVE has two parts: (1) the visual tokens from the Video LLM's visual encoder $\mathbf{z}^v$, extracted at sparsely sample video key frames, and (2) the

side-channel visual tokens $\mathbf{z}^s$, derived from a high frame rate video.

Visual Encoder. For $\mathbf{z}^v$, we use the same setting as the one in Section B.2.

Side-Channel Signal Encoder. In this case, the side-channel signals $\mathbf{z}^s$ are high frame rate videos. We sample the video frames at the frame rate of 2fps and use the default pre-processing step of the LanguageBind to reshape and crop the video frames. To leverage LanguageBind [23] to encode the high-frame-rate video frames, we split the video frames along the temporal axis into multiple non-overlap groups with each group containing 8 frames. We later concatenate the encoded features of all groups along the temporal axis. To reduce the overhead of the PAVE, inspired by the Slow-Fast [5], we downsample the spatial resolution of the video feature of each video frame from 16×16 to 2×2. We do not utilize the classification tokens from the output of the LanguageBind. Since we do not fine-tune LanguageBind's video encoder, we pre-extract the video features in order to speed up the training.

Network Architecture and Training Details. For the PAVE design and the training configuration, we use the same hyper-parameters used in Section B.2.

Training data. We create a subset from LLaVA-Video-178K [21] by first sampling all videos longer than 1 minute and then randomly choosing 2 question-answer pairs for each video. This process creates a training set that contains 57K videos and 114K question-and-answer pairs.

Evaluation benchmark. We use VideoMME [6], MVBench [11], and MLVU [22] as evaluation benchmarks. VideoMME and MVBench are both comprehensive video benchmarks and cover different types of subtasks, while MLVU focuses on long video understanding. VideoMME includes 6 key domains and 30 sub-classes. It contains 900 videos, ranging from less than one minute to nearly one hour. There are 2,700 questions with each accompanied by four options. MVBench includes 20 different sub-tasks, such as object shuffling and fine-grained pose estimation, which require detailed temporal information. In total, it has about 4000 questions and 3900 videos. MLVU contains 2175 questions and 1337 long videos. All benchmarks adopt accuracy as the performance metric.

B.5. The Calculation of Inference FLOPs.

We now describe how the floating-point operations (FLOPs) are reported in our experiments. Since the visual-encoder and the side-channel information encoder are replaceable modules in PAVE settings (i.e. we can use encoder with different scales at different settings.), we only consider the FLOPs of PAVE and LLM, provided by the LLM-Viewer [19]. During the FLOPs calculation of LLM, we consider 6272 visual tokens, and following the previous work [16], we add 40 additional tokens for the text. We then calculate and add the FLOPs of PAVE.

The FLOPs of PAVE is calculated as follows. The input of the PAVE consists of two parts, the visual tokens $\mathbf{z}^v$ from the Video LLM's visual backbone, and the side-channel information tokens $\mathbf{z}^s$. We consider the case that the visual tokens $\mathbf{z}^v$ come from 32 video frames and Video LLM's visual backbone generates 196 tokens for each frame.

- **Audio-visual QA**: We assume the length of the video at inference time is 2 minutes and the audio encoder will generate 1 token for each second of the audio. It yields 120 audio tokens. The cross-attention is conducted over 196 query tokens and 4 key tokens. PAVE thus introduces about 0.07 TB and 0.10 TB FLOPs for 0.5B and 7B models, respectively.
- **3D QA**: We uniformly sample 32 frames and send them into the 3D backbone. It generates 576 tokens for each frame and yields 18432 tokens in total. The cross-attention is conducted over 196 query tokens and 576 key tokens. PAVE introduces about 0.12 TB and 0.15 TB FLOPs for 0.5B and 7B models, respectively.
- **Enhancing video QA**: We assume the length of the video at inference time is 2 minutes—close to the average duration of videos on VideoMME and MVBench. We sample the frames at 2 fps and sent them to the video backbone. We down-sample the tokens of each frame spatially to 2 by 2 grids. It produces 960 video tokens in total. The cross-attention is conducted over 196 query tokens and 30 key tokens. PAVE adds about 0.07 TB and 0.10 TB FLOPs for 0.5B and 7B models, respectively.
- **Multi-view Video Understanding**: We uniformly sample 32 frames for each exo-centric video and send them into the SigLIP. It generates 196 tokens for each frame and yields 25,088 tokens for 4 exo-centric videos in total. The cross-attention is conducted over 196 query tokens and 784 key tokens. PAVE adds about 0.14 TB and 0.17 TB FLOPs for 0.5B and 7B models, respectively.

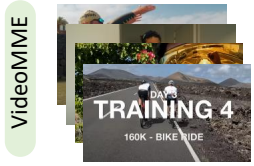
C. Additional Visualization

We present additional visualization of the PAVE's results for enhanced video QA in Figure 1 with videos from VideoMME [6] and MVBench [11].

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 2, 3

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial



VideoMME

Question

Which of the following options does not match the description in the video?
A. On the first day of training, the two male protagonists ran 8km and rested most of the rest of the time.
B. On the second day of training, the two male protagonists swam three kilometers in open water and then cycled 60 kilometers.
C. On the third day of training, the two male protagonists rode a total of 160km, after which only one of them proceeded to run an additional 4km.
D. On the last day of training, the two male protagonists conducted a simulated triathlon training, swimming 3000m, cycling 40km, and running 16km. Both of them successfully completed.

LLaVA-OV (zero-shot): A

PAVE: C

GT: C

VideoMME

Question

According to the content of the video, why does the little boy squatting on the ground playing with a toy car run away and hide?
A. Because he hears a loud thunderclap, mistaking it for an explosion, and becomes terrified
B. Because he sees several gangsters walking towards him and his mother with big knives, he is very scared.“
C. Because he sees soldiers pointing guns at him and his mother, and he is very frightened.
D. Because he accidentally broke the toy car and doesn't want his mother to find out, fearing disappointment rather than punishment.

LLaVA-OV (zero-shot): A

PAVE: C

GT: C

MVBench

Question

The person uses multiple similar objects to play an occlusion game. Where is the hidden object at the end of the game from the person's point of view?

Candidates:
Under the first object from the left.
Under the third object from the left.
Under the second object from the left.

LLaVA-OV (zero-shot): Under the first object from the left.

PAVE: Under the third object from the left.

GT: Under the third object from the left.

Figure 1. Visualization of the QA results on enhanced video QA task. By making use of the video feature of the densely sampled video frames, PAVE captures more details in the video and thus improves the performance of video understanding.

scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19129–19139, 2022. 2, 3

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 3

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4

[6] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren,

Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 4

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 2

[8] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1, 3

[9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 3

[10] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[11] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 4

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 3

[14] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. 2, 3

[15] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video dialog as conversation about objects living in space-time. In *European Conference on Computer Vision*, pages 710–726. Springer, 2022. 3

[16] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 4

[17] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. 2, 3

[18] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer, 2024. 3

[19] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, et al. LLM inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*, 2024. 4

[20] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 2

[21] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 4

[22] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2025. 4

[23] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[24] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 3