

OneLLM: One Framework to Align All Modalities with Language

Supplementary Material

Jiaming Han^{1,2}, Kaixiong Gong^{1,2}, Yiyuan Zhang^{1,2}, Jiaqi Wang², Kaipeng Zhang²
Dahua Lin^{1,2}, Yu Qiao², Peng Gao², Xiangyu Yue^{1†}

¹MMLab, The Chinese University of Hong Kong
²Shanghai Artificial Intelligence Laboratory

A. Appendix Overview

- Sec. **B**: Additional Ablation Experiments.
- Sec. **C**: Additional Implementation Details.
- Sec. **D**: Evaluation Details.
- Sec. **E**: Comparison with Prior Works.
- Sec. **F**: Additional Qualitative Results.

B. Additional Ablation Experiments

Encoder Type	Frozen	Mem.	Nocaps	VQAv2	ClothoQA	MSVDQA
CLIP	✓	46Gb	115.8	71.6	57.4	56.8
CLIP	✗	74Gb	106.0(-9.8)	69.1(-2.5)	62.1(+4.7)	53.6(-3.2)
DINOv2	✓	33Gb	104.6(-11.2)	67.0(-4.6)	56.8(-0.6)	54.7(-2.1)

Table 1. Ablation Experiments on Universal Encoder.

In the main paper, we follow previous works [52] and set a frozen CLIP-ViT as the universal encoder. Here we explore other design choices such as trainable CLIP-ViT and DINOv2 [35] as the encoder.

Frozen vs. Trainable Encoder. We first turn on all the parameters in the multimodal-text alignment stage. As shown in Tab. 1, the performance for visual modalities (image and video) dropped significantly, while the result for audio QA (ClothoQA) improved by 4.7%. We think trainable CLIP will break the pretrained vision-language representations but can leave more space for learning other modalities. However, considering the memory usage (46Gb vs. 74Gb), frozen CLIP will be a better choice for our framework.

Beyond Vision-Language Encoder. In addition to the vision-language encoder CLIP-ViT, we also explore other models, such as the self-supervised vision model DINOv2 [35], as the universal encoder. In Tab. 1, we noticed that the performance of OneLLM using DINOv2 is lower than the model using CLIP-ViT because DINOv2 is

not aligned with language and we need to learn the vision-language alignment from scratch.

C. Additional Implementation Details

C.1. Lightweight Modality Tokenizers

The modality tokenizer is to transform input signal into a sequence of tokens. Here we will introduce the tokenizer of each modality in detail.

Visual Tokenizer. We use the same tokenizer setting for visual modalities, *i.e.*, image, video, depth/normal map. The visual tokenizer is a single 2D convolution layer:

$$\text{Conv2D}(C_{in} = 3, C_{out} = 1024, K = (14, 14), S = (14, 14)), \quad (1)$$

where C_{in} , C_{out} , K and S denote the input channel, output channel, kernel size and stride, respectively. Note that for a video input $\mathbf{x} \in \mathbb{R}^{T \times H \times W}$ with T frames, height H and width W , we parallel feed its frames into the tokenizer, resulting in $T \times \frac{H}{14} \times \frac{W}{14}$ tokens. Similarly, image, depth/normal map can also be regarded as a one-frame video input $\mathbf{x} \in \mathbb{R}^{1 \times H \times W}$.

Audio Tokenizer. We first transform audio signals into 2D spectrogram features $\mathbf{x} \in \mathbb{R}^{1 \times H \times W}$, where $H=128$ and $W=1024$ by default. Following [13], the audio tokenizer is a single 2D convolution layer:

$$\text{Conv2D}(C_{in} = 1, C_{out} = 1024, K = (16, 16), S = (10, 10)). \quad (2)$$

Point Tokenizer. For a raw point cloud, we sample 8192 points using Furthest Point Sampling (FPS), resulting in a 2D tensor $\mathbf{x} \in \mathbb{R}^{8192 \times 6}$. Then we use the KNN algorithm to group these points into 512 groups: $\mathbf{x} \in \mathbb{R}^{512 \times 32 \times 6}$ where 32 is the size of each group. After that, we encode the point cloud with a 2D convolution layer:

$$\text{Conv2D}(C_{in} = 6, C_{out} = 1024, K = (1, 1), S = (1, 1)), \quad (3)$$

followed by a `max` operation on dimension 1. Finally, the shape of output tokens is $\mathbb{R}^{1024 \times 1024}$.

IMU Tokenizer. For an IMU input with shape $\mathbb{R}^{2000 \times 6}$, we tokenize it with a 1D convolution layer:

$$\text{Conv1D}(C_{in} = 6, C_{out} = 1024, K = 10, S = 1), \quad (4)$$

resulting in a sequence of tokens $\mathbf{x} \in \mathbb{R}^{1024 \times 391}$.

fMRI Tokenizer. The shape of an fMRI signal is \mathbb{R}^{15724} . We tokenize it with a 1D convolution layer:

$$\text{Conv1D}(C_{in} = 15724, C_{out} = 8196, K = 1, S = 1). \quad (5)$$

We then resize the output tensor $\mathbf{x} \in \mathbb{R}^{8196}$ into a 2D tensor $\mathbf{x} \in \mathbb{R}^{1024 \times 8}$ to align with the input of the transformer encoder.

C.2. Multimodal-Text Alignment Dataset

We summarize the multimodal-text alignment dataset in Tab. 2. For depth/normal-text pairs, we adopt DPT model [37] pretrained on omnidata [11] to generate depth/normal map. The source dataset is a subset of CC3M [42], around 0.5M image-text pairs. For IMU-text pairs, we use the IMU sensor data of Ego4D [15] and the corresponding video narrations (*i.e.*, text annotations). For fMRI-text pairs, we use the subj01 imaging session of NSD [4] and follow the same data split with [41]. Note that the visual stimulus, *i.e.*, images shown to participants, are from MS COCO [8]. Therefore, we use the image captions in COCO Captions as text annotations of fMRI-text pairs.

Modality	Multimodal-Text Alignment		Multimodal Instruction Tuning	
	Size	Dataset	Size	Dataset
Image	1000M	LAION-400M [39] LAION-COCO [38]	1216K	LLaVA-150K [26] COCO Caption [8] VQAv2 [14], GQA [18] OKVQA [30], A-OKVQA [40] OCRvQA [32], RefCOCO [19] Visual Genome [21]
Video	2.5M	WebVid-2.5M [5]	461K	MSRVTT-Cap [50] MSRVTT-QA [49] Video Conversation [53]
Audio	0.4M	WavCaps [31]	60K	AudioCaps [20] Audio Conversation [53]
Point	0.6M	Cap3D [29]	70K	Point Conversation [51]
Depth	0.5M	CC3M [42]	50K	LLaVA-150K [26]
Normal	0.5M	CC3M [42]	50K	LLaVA-150K [26]
IMU	0.5M	Ego4D [15]	50K	Ego4D [15]
fMRI	9K	NSD [4]	9K	NSD [4]
Text	-	-	40K	ShareGPT [1]
Total	1005M		2006K	-

Table 2. Training Datasets.

C.3. Multimodal Instruction Tuning Dataset

We summarize the multimodal instruction tuning dataset in Tab. 2.

C.4. Prompt Design

The prompt formats for each dataset are shown in Tab. 3.

Dataset	Prompt Format
LLaVA-150K [26] ShareGPT [1] Video Conversation [53] Audio Conversation [53] Point Conversation [51]	(use their original prompt)
VQAv2 [14], GQA [18] OKVQA [40] OCRvQA [32] MSRVTT-QA [49]	{Question} Answer the question using a single word or phrase.
A-OKVQA [40]	{Question} {Options} Answer with the option's letter from the given choices directly
TextCaps [43] COCO Caption [8] MSRVTT-Cap [50] AudioCaps [20]	Provide a one-sentence caption for the provided image/video/audio.
RefCOCO [19] Visual Genome [21]	Provide a short description for this region.
Ego4D [15]	Describe the motion.
NSD [4]	Describe the scene based on fMRI data.

Table 3. Prompt Formats for Training.

D. Evaluation Details

Dataset	Prompt Format
MMVet	(use the original prompt)
GQA [18] VQAv2 [14] OKVQA [30] TextVQA [44] MME [12] MSVD [49] Clotho AQA [25] MUSIC-AVQA [23]	{Question} Answer the question using a single word or phrase.
ScienceQA [28] MMBench [27] SEED-Bench [22] NextQA [48] How2QA [24]	{Question} {Options} Answer with the option's letter from the given choices directly
VizWiz [16]	{Question} When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
Nocaps [2] Flickr30K [36] VATEX [47] VALOR [7] Clotho Cap [10] Objaverse-Cap [9]	Provide a one-sentence caption for the provided image/video/audio/point cloud.
AVSD [3]	{Question} Answer the question and explain the reason in one sentence.
Objaverse-CLS [9]	What is this?
NYUv2 [34] SUN RGB-D [45]	{Class List} What is the category of this scene? Choice one class from the class sets.

Table 4. Prompt Formats for Evaluation.

In this section, we first list the evaluation prompts for each dataset in Tab. 4. Then we will give more evaluation details.

Model	Encoder Param	#Encoder	#Projection	Supported Modalities								
				Image	Video	Audio	Point	IMU	Depth	Normal	fMRI	
X-LLM [6]	-	3	3	✓	✓	✓						
PandaGPT [46]	1.2B	2	1	✓	✓	✓						
ImageBind-LLM [17]	1.8B	3	1	✓	✓	✓	✓					
ChatBridge [53]	1.3B	3	3	✓	✓	✓						
AnyMAL [33]	2B	3	3	✓	✓	✓			✓			
OneLLM (Ours)	0.6B	1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 5. Comparisons of Different Multimodal LLMs.

Image, Video and Audio Tasks. We evaluate all datasets using their official evaluation protocols. As shown in Tab. 4, for QA tasks with options, we ask OneLLM to directly predict the option letters; For open-ended QA tasks, we ask OneLLM to predict a single word or phrase. For captioning tasks, we ask OneLLM to generate a one-sentence caption. Note that for audio-video-text tasks, the input sequence to the LLM is: $\{Video\ Tokens\} \{Audio\ Tokens\} \{Text\ Prompts\}$.

Point Cloud Tasks. Our evaluation on point cloud tasks mainly follows PointLLM [51]. For the point cloud classification task, we use the same prompt as PointLLM: *What is this*, and evaluate the accuracy using GPT4.

Depth/Normal Map Tasks. For scene classification using depth/normal map, we first prepend the category list to the beginning of prompt, then we ask OneLLM to choose one class for the list.

IMU/fMRI Tasks. We evaluate on IMU/fMRI captioning tasks. The prompts are the same as their training prompts: *Describe the motion* for IMU captioning and *Describe the scene based on fMRI data* for fMRI captioning.

E. Comparison with Prior Works

The main difference between OneLLM and previous MLLMs is that we show a unified encoder is *sufficient* to align multi-modalities with LLMs. As shown in Tab. 5, OneLLM with *one* universal encoder, *one* projection module and *less* parameters (0.6B) can unify more modalities into one framework. The results in the main paper (Tab. 1-6) also demonstrate that OneLLM can achieve better performance to previous works. The ablation experiments in Tab. 7 (a) also show that jointly training all modalities with our unified framework can benefit data-scarce modalities. Here we are not trying to prove that OneLLM’s architecture is optimal, but to show the possibility of building MLLMs using a unified and scalable framework.

F. Additional Qualitative Results

In this section, we provide more qualitative results in Fig. 1, Fig. 2 and Fig. 3.

References


- [1] Sharegpt. <https://sharegpt.com/>, 2023. 2
- [2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocos: novel object captioning at scale. In *ICCV*, 2019. 2
- [3] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio-visual scene-aware dialog. In *CVPR*, 2019. 2
- [4] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 2
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2
- [6] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 3
- [7] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weinong Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 2
- [8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2
- [9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 2
- [10] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, pages 736–740. IEEE, 2020. 2

What do you think of the photo?



The photo is a captivating and dynamic scene of a bustling city street in New York City. The image features a large number of billboards and advertisements on the buildings surrounding the street, creating a vibrant atmosphere. The street is filled with people walking and cars driving, giving the impression of a busy urban environment. The photo is taken from a low angle, which adds a sense of depth and perspective to the scene. The traffic lights are visible, indicating that the street is well-regulated for pedestrian and vehicular safety. The overall composition of the photo is visually striking, showcasing the energy and excitement of the city.

Where might this picture have been taken?



This picture might have been taken in a large city with a lot of tall buildings. The presence of a red double-decker bus and a truck in the image suggests that the city has a public transportation system and possibly a busy traffic environment. The tall buildings and the presence of a bus and truck indicate that the city is likely a bustling urban area with a high population density.

Choose one from "Sydney", "New York" and "Hong Kong", and tell me the reason.


The presence of a red double-decker bus and a truck in the image suggests that the city could be either Hong Kong or New York.

Why this image is unusual?




The image is unusual because it features a dog dressed in a chef's outfit, complete with a hat, apron, and a bow tie. This is not a typical sight, as dogs are not typically dressed in human clothing or accessories. The dog is sitting at a table with a bowl of food, which adds to the unconventional nature of the scene. The image is likely meant to be humorous or creative, as it is not a common occurrence to see a dog dressed in human attire.

Tell me about this place.



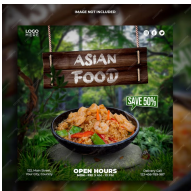
The image features a large, grassy field with a group of stone structures, possibly Stonehenge, situated in the center. The structures are made of rocks and are arranged in a circle, with some rocks appearing to be taller than others. The sky above the field is filled with clouds, creating a beautiful and serene atmosphere. The scene is reminiscent of ancient civilizations and their architectural achievements.

What does this picture want to express?




The image is a symbolic representation of the bond between a parent and their child. The hand of an adult is holding a small plant, which is a symbol of growth and life, while the hand of a baby is reaching out to the plant. This image suggests the nurturing and care that parents provide to their children, as well as the innocence and curiosity of a child. The scene is set in a grassy field, which adds to the natural and serene atmosphere of the image.

Introduce me this poster.




The poster features a delicious Asian food advertisement, with a large bowl of rice and shrimp displayed prominently. The advertisement is designed to entice customers with its mouth-watering presentation. The poster is displayed in a restaurant, with a sign that reads "Open Hours" and "50% Off" in the background. This suggests that the restaurant is offering a special deal on Asian food during certain hours.

Why this image is funny?



This image is funny because it features three raccoons dressed in human clothes and driving a car. The raccoons are wearing sunglasses and one of them is playing a guitar, which adds a humorous and unexpected element to the scene. The fact that raccoons are not capable of driving or playing instruments makes the situation comical and amusing to viewers.

If you could only use three words to express the theme of this picture, what would they be?




Foggy, tree, person

Provide a short caption for this image

A person running in a field with a tree in the background.

What iconic attractions appear in this picture?



The picture features several iconic attractions, including the Statue of Liberty, the Eiffel Tower, the Roman Coliseum, the pyramids of Egypt, the Space Needle, and the London Bridge.

Figure 1. Additional Qualitative Image Demos.








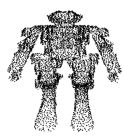



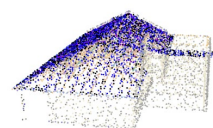
<p>Describe the video in detail</p>  <p>A man is playing basketball in a gym. He is wearing a black shirt and white pants. He is bouncing a basketball on his finger and then dribbling it with his hand. He is also wearing a black and white striped shirt. The gym is white and has a white floor. There is a white wall behind him.</p>	<p>What activity is going on in the video?</p>  <p>The video shows a rocket launching from a launch pad.</p> <p>Where might this launch pad be located?</p> <p>It is not clear from the video where the launch pad is located. It could be in a spaceport or a military base.</p>	<p>What process does this video record?</p>  <p>This video records the process of a flower blooming.</p> <p>Can you provide more details?</p> <p>The flower is white and has a long stem. The flower is in a greenhouse, and the camera is recording the flower's growth process. The flower is in bloom, and the camera captures the flower's growth and development.</p>
<p>What can you hear in the audio?</p>  <p>(bell ringing)</p> <p>I can hear the sound of a bell ringing at a steady pace.</p> <p>Can you provide more details?</p> <p>The bell is large and metallic, and it is ringing in a melody that is loud and clear.</p>	<p>Where could this audio take place?</p>  <p>(bird singing)</p> <p>The audio could take place in a place with birds, like a forest or a park.</p> <p>What kind of bird call could this be?</p> <p>It could be a bird call like a whistle or a chirp, but without more context, it's hard to say for sure.</p>	<p>Describe the audio in detail</p>  <p>(crunching sound)</p> <p>In the audio, a person is seen shuffling through a variety of items, and the items make a crunching sound. The person then takes a moment to speak. The items in the background are not clear, but they appear to be the source of the crunching sound.</p>
<p>Summary the content of the audio briefly.</p>  <p>(speaking and laughing)</p> <p>A woman speaking followed by a baby laughing</p> <p>Describe the audio in detail</p> <p>In the audio, a woman is speaking to a laughing baby. The woman's voice is calm and soothing, while the baby's laughter is contagious. It sounds like they are having a conversation, with the woman saying something that makes the baby laugh even more. The audio also contains some background noise, but it is not clear what it is.</p>	<p>What is this?</p>  <p>This is a 3D model of a cartoon-style humanoid figure, predominantly white in color. The figure is designed in a simplistic, stylized manner, typical of cartoon characters. It has a neutral expression, suggesting it could be used in a variety of contexts, such as animation, game development, or graphic design. The figure's white color allows for easy customization, such as adding clothing or other accessories.</p>	<p>How would you summarize this 3D data set?</p>  <p>This 3D data set represents a toy car, predominantly blue in color, with a distinctive white stripe running along its length. The car features a black windscreen and matching black windows, adding to its realistic appearance. The toy car, with its vibrant colors and detailed design, is likely to be a popular choice among children for play and could also serve as a decorative item.</p>
<p>What kind of object is illustrated by this collection of points?</p>  <p>This is a 3D model of a cartoon-style dog, predominantly in a vibrant yellow color. The model is characterized by its exaggerated features, typical of cartoon aesthetics, such as large, expressive eyes and a broad, friendly smile. The dog's design suggests it could be used in animations, video games, or children's educational content.</p>	<p>Provide an outline of this 3D shape's characteristics.</p>  <p>The 3D object model is a tree with a distinctive green color. It features a sturdy trunk and a dense canopy of leaves, giving it a realistic appearance. The tree's design suggests it could be used in various digital environments, such as video games, animations, or virtual reality simulations.</p>	<p>Describe the 3D model in detail</p>  <p>The 3D model is a cartoon-style house with a vibrant purple roof. The house is simplistic in design, with a single story and a single room. The roof is a striking purple color, which contrasts with the white walls. The model is likely used in animation or game design, where such whimsical, colorful elements are common.</p>

Figure 2. Additional Qualitative Video, Audio and Point Cloud Demos.

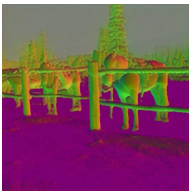
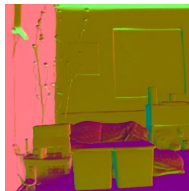
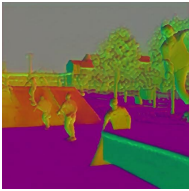
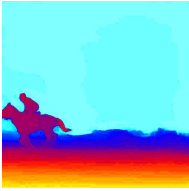

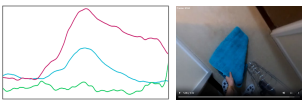
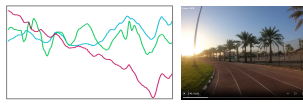
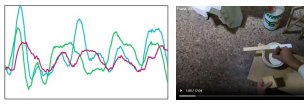



<p>Describe the scene based on the normal map</p>  <p>The normal map shows a group of horses standing in a fenced-in area, possibly a corral. There are five horses in total, with four of them standing close together, and one horse standing slightly apart from the others. The horses are of various sizes and are positioned in a way that suggests they are waiting or resting. The fence surrounding the area is made of wood, and the ground is covered with dirt.</p>	<p>What's in this room?</p>  <p>The room is dimly lit, with a bed in the center and a small wooden table nearby. There is also a chair in the room. The bed is made up with a white comforter and pillows, and there is a lamp on the table. The room is decorated with several potted plants, including a large one near the bed. Additionally, there is a vase with flowers in it, and a clock on the wall. The overall atmosphere of the room is cozy and inviting.</p>	<p>What are people doing in the normal map?</p>  <p>In the normal map, people are skateboarding and performing tricks at a skate park.</p>
<p>What can you know from the depth map?</p>  <p>The depth image shows a man riding a horse through a snowy field, with the horse's tail and mane blowing in the wind. The man is wearing a black jacket, and the horse is a dark brown color. The scene is set against a backdrop of a sunset, creating a beautiful and serene atmosphere. The depth image captures the essence of the man and his horse enjoying the outdoors during the winter season.</p>	<p>What should you do in this situation?</p>  <p>In this situation, it is important to remain calm and cautious. Since the bear is walking towards the camera, it is essential to avoid sudden movements or loud noises that might startle the bear. If the bear approaches, it is crucial to maintain eye contact with the bear to show that you are aware of its presence and not a threat. If the bear starts to charge, it is advisable to back away slowly and carefully, keeping an eye on the bear to ensure it does not change direction.</p>	<p>What are the two people doing?</p>  <p>The two people are playing baseball, with one person holding a bat and the other person holding a glove.</p>
<p>What happened in the motion?</p>  <p>(GT: picks the mats) (reference video)</p> <p>I can see a person walking on the floor and picking something up.</p>	<p>What is the person doing?</p>  <p>(GT: walks in the stadium) (reference video)</p> <p>The person is walking on the sidewalk.</p>	<p>Describe the motion.</p>  <p>(GT: dips the brush in the paint) (reference video)</p> <p>The motion is like a dance, with the paintbrush moving in a smooth and controlled motion.</p>
<p>What scene does this fMRI data record?</p>  <p>(reference visual stimulus)</p> <p>A man riding a kiteboard on a long wave in the ocean.</p>	<p>What is shown in the fMRI data?</p>  <p>(reference visual stimulus)</p> <p>The fMRI data show the brain activity about a zebra.</p>	<p>What scene does this fMRI data record?</p>  <p>(reference visual stimulus)</p> <p>The fMRI data record the scene of a flower garden.</p>

Figure 3. Additional Qualitative Depth/Normal Map, IMU and fMRI Demos.

- [11] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 2
- [12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2
- [13] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 1
- [14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 2
- [15] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 2
- [16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 2
- [17] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 3
- [18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2
- [19] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2
- [20] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 2
- [21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 2
- [22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2
- [23] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Jirong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118, 2022. 2
- [24] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 2
- [25] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowd-sourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE, 2022. 2
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 2
- [27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2
- [28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 2
- [29] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 2
- [30] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2
- [31] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023. 2
- [32] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE, 2019. 2
- [33] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Any-mal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023. 3
- [34] Pushmeet Kohli, Nathan Silberman, Derek Hoiem, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. 2
- [35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1
- [36] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 2
- [37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 2
- [38] C. Schuhmann, A. Köpf, R. Vencu, T. Coombes, and R. Beaumont. Laion coco: 600m synthetic captions from laion2b-en. 2

- [39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 2
- [40] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162. Springer, 2022. 2
- [41] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind’s eye: fmri-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023. 2
- [42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 2
- [43] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758. Springer, 2020. 2
- [44] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 2
- [45] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 2
- [46] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 3
- [47] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatec: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019. 2
- [48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 2
- [49] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 2
- [50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 2
- [51] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointlm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 2, 3
- [52] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 1
- [53] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: