

AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model

Seungwhan Moon* Andrea Madotto* Zhaojiang Lin* Tushar Nagarajan*
 Matt Smith Shashank Jain Chun-Fu Yeh Prakash Murugesan
 Peyman Heidari Yue Liu Kavya Srinet Babak Damavandi Anuj Kumar
 FAIR, Meta & Meta Reality Labs

Abstract

We present Any-Modality Augmented Language Model (AnyMAL), a unified model that reasons over diverse input modality signals (*i.e.* text, image, video, audio, IMU motion sensor), and generates textual responses. AnyMAL inherits the powerful text-based reasoning abilities of the state-of-the-art LLMs including LLaMA-2 (70B), and converts modality-specific signals to the joint textual space through a pre-trained aligner module. To further strengthen the multimodal LLM’s capabilities, we fine-tune the model with a multimodal instruction set manually collected to cover diverse topics and tasks beyond simple QAs. We conduct comprehensive empirical analysis comprising both human and automatic evaluations, and demonstrate state-of-the-art performance on various multimodal tasks.

1 Introduction

Large Language Models (LLMs), known for their substantial size and complexity, have significantly enhanced the capacity of machines to understand and articulate human language. The progress in LLMs has also led to notable advancements in the vision-language domain [1, 2, 3, 4], bridging the gap between image encoders and LLMs to combine their reasoning capabilities. Prior multimodal LLM research has concentrated on models that combine text and one other modality [3, 5], such as text and image models, or has centered on proprietary language models that are not open sourced [2, 4].

To tackle the previously mentioned challenges, we introduce **Any-Modality Augmented Language Model** (AnyMAL) — a collection of multi-modal encoders trained to transform data from various modalities, including images, videos, audio, and IMU motion sensor data, into the text embedding space of an LLM. To achieve this, we extend the work by [1] to (1) more capable instruction-tuned LLMs (*i.e.* LLaMA-2-70B-chat [6]), (2) larger pre-trained modality encoders, and (3) advanced projection layers to handle variable input lengths. The model output examples are shown in Figure 1, and an illustration of the overall methodology is shown in Figure 2.

The key contributions of the work are as follows:

- We present an efficient and scalable solution for building Multimodal LLMs. We provide projection layers pre-trained on large datasets with diverse modalities (*e.g.* 200M images, 2.2M audio, 500K IMU time-series, 28M videos) all aligned to the same LLM (LLaMA-2-70B-chat), thus enabling interleaved multimodal in-context prompting.
- We further fine-tune the model with the multimodal instruction set across three modalities (image, video, and audio) covering diverse unconstrained tasks beyond simple QA domains. The dataset features high-quality manually collected instruction data, which we thus also use as a benchmark for complex multimodal reasoning tasks.
- Our best model achieves strong zero-shot performance in both automatic and human evaluation on diverse tasks and modalities, setting new SOTA with +7.0% relative accuracy

*Joint First Authors. ✉: {shanemoon, andreamad8, zhaojiang, tusharn}@meta.com

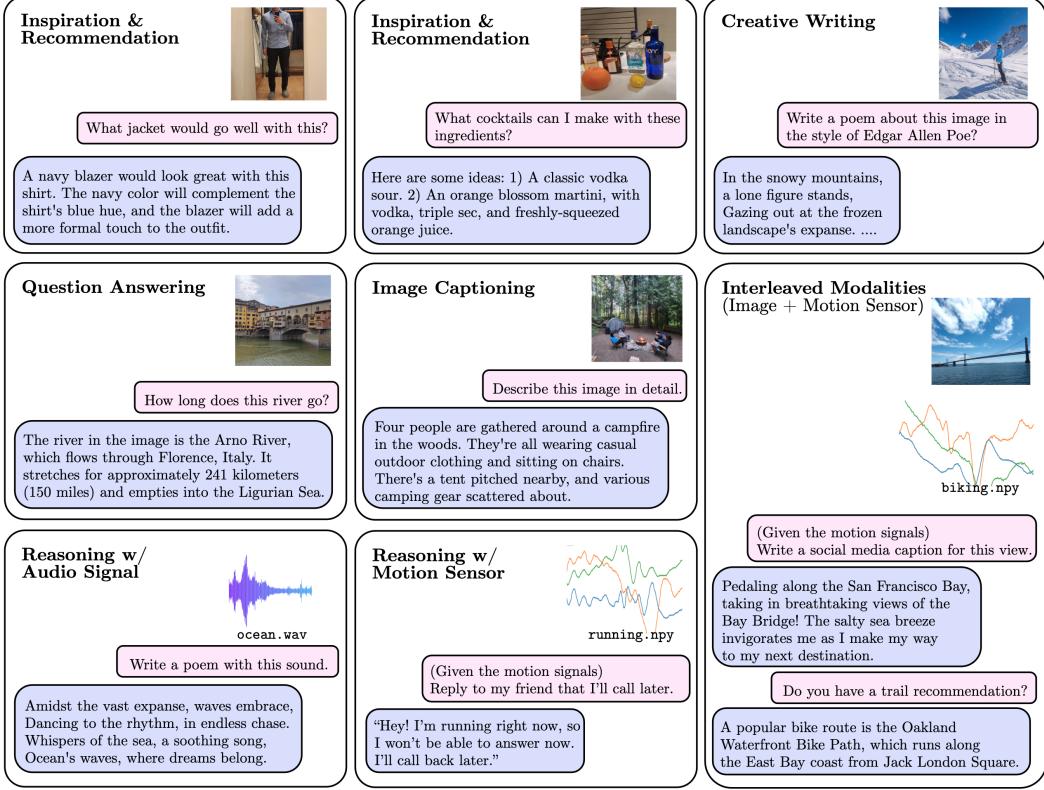


Figure 1: Example AnyMAL outputs. The model understands various input signals (*i.e.* vision, audio, motion sensor signals), and responds to free-form user queries. When multiple modalities are interleaved and given as input (*e.g.* right-most: image + IMU motion sensor signals), the model reasons over them jointly.

improvement on VQAv2, +8.4% CIDEr on zero-shot COCO image captioning, and +14.5% CIDEr on AudioCaps, when compared with the models available in the literature.

2 Related Work

Large Language Models (LLM): There has been a surge of LLMs with varying model sizes recently, showcasing remarkable reasoning capabilities. While the most well-known commercial service is ChatGPT [4, 7], the open-sourced models include FlanT5 [8], GPT-J [9], OPT [10], LLaMA [11], Vicuna [12], and more recently, LLaMA-2 [6].

Our work builds upon the powerful text-based reasoning capabilities of these LLMs, extending these capabilities to multimodal inputs.

Vision-Language Models: Numerous studies have addressed the task of instructing a unified model that integrates both visual and linguistic elements, finding practical implementations in domains like image captioning [13] and visual question answering (VQA) tasks [14, 15, 16]. While the relative scarcity of data sources aligning different modalities has conventionally been considered the bottleneck in scaling, recent works have shifted towards harnessing the capabilities of pre-trained LLMs, tapping into the knowledge accrued from extensive textual corpora. These work include Flamingo [2], OpenFlamingo [17], Palm-E [18], BLIP-2 [3], InstructBLIP [19], LLaVA [20], IDEFICS [5], MiniGPT-4 [21] and many more [22, 23, 24, 25, 26, 27, 28], where each model uses different variants of base LLMs. These models typically undergo fine-tuning stages as well, re-purposing several task-specific vision-language datasets [20, 29].

Our work extends the previous approaches by (1) allowing for diverse input modalities beyond vision signals, (2) presenting a fine-tuning process with our manually collected multimodal instruction tuning data, and (3) scaling the LLM parameters to 70B via an efficient pre-training approach.

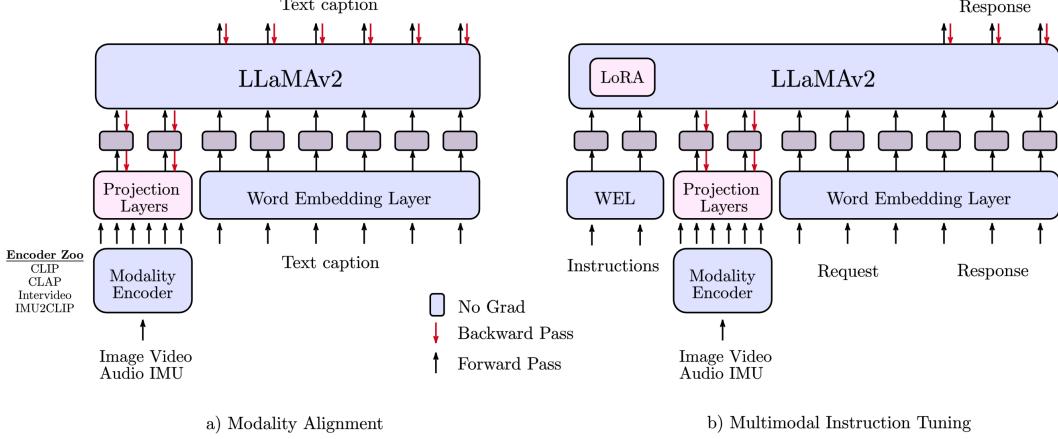


Figure 2: AnyMAL Training. (a) Modality alignment pre-training allows for mapping the output of each modality encoder into the joint LLM embeddings space through projection layers. (b) With multimodal instruction tuning, the model learns to associate system instructions and text queries with input multimodal contexts. Our modality-specific encoder zoo includes: CLIP ViT-L, ViT-G, DinoV2 (image), CLAP (audio), IMU2CLIP (IMU motion sensor), and Intervideo (video).

3 Methods

3.1 Pre-training

Modality Alignment: We achieve the multimodal understanding capabilities by pre-training LLMs with paired multimodal data (modality-specific signals and text narrations) (Figure 2). Specifically, we train a lightweight adapter for each modality to project the input signals into the text token embedding space of a specific LLM. In this way, the text token embedding space of the LLM becomes a joint token embedding space, with tokens representing either text or other modalities. The number of token embeddings used to represent each input modality is fixed per adapter, ranging from 64 - 256 in this work. During the alignment training, we freeze the model parameters of the underlying LLM, which allows it to reach convergence faster than training end-to-end from scratch, and to inherit the reasoning capabilities of the LLM at inference time. In addition, to maximize the feature compatibility, for each modality we use an encoder $g(\cdot)$ that has already been aligned to a text embeddings space, *e.g.* CLIP [30, 31] for images, CLAP [32] for Audio signals, or IMU2CLIP [33] for IMU signals. For each text caption and modality pair $(\mathbf{X}_{\text{text}}, \mathbf{X}_{\text{modality}})$, we align them using the following objectives with a projection module (*i.e.* Perceiver Resampler [2] for vision encoder, and linear layers for other modalities).

$$p(\mathbf{X}_{\text{text}} | \mathbf{X}_{\text{modality}}) = \prod_{i=1}^L p_\theta(\mathbf{X}_{\text{text}}^{[i]} | \mathbf{Z}_{\text{modality}}, \mathbf{Z}_{\text{text}}^{[1:i-1]}) \quad (1)$$

$$\mathbf{Z}_{\text{modality}} = \text{Projection}_\theta(h_{\text{latents}}, g(\mathbf{X}_{\text{modality}})) \quad (2)$$

Datasets: For image alignment, we use a cleaned subset of the LAION-2B dataset, filtered using the CAT method and with any detectable faces blurred [34]. For audio alignment, we use AudioSet [35] (2.1M samples), AudioCaps [36] (46K samples), and CLOTHO [37] (5K samples) datasets. We use the Ego4D dataset [38] for IMU and text alignment (528K).

Quantization: Scaling the pre-training to 70B parameter models for a large dataset (200M+ instances) requires significant resources, often requiring an FSDP [39] wrapper to shard the model across multiple GPUs. To effectively scale our training, we implement the quantization strategies (4 bits and 8 bits) [40] in our multimodal settings, in which we keep the LLM component of our model frozen and only have the modality tokenizers trainable. This approach shrinks the memory requirement by an order of magnitude. Thus, we are able to train 70B AnyMAL on a single 80GB VRAM GPU with a batch size of 4.

Compared with FSDP, we observe that the proposed quantization approach achieves the same throughput while only using half of the GPU resources. We do note that the training / validation loss were constantly higher compared to the FSDP training, but nevertheless did not impact the generation quality (at inference time, we use the original LLM at full precision to maximize the accuracy).

Example 1. Creative Instructions (Image)

**Instruction Prompt**

Write a short story about the scene in this photo. Include some dialogue between the characters.

Response

The restaurant table stood empty but adorned with remnants of a meal. A seagull landed on its edge, spreading his wings in a display of dominance. With a raucous cry, he fiercely guarded the plate before him, which he intended to make his dinner. His piercing gaze fixed upon another seagull, wings poised for landing, his eyes locked on the feast. "Back off!" the first seagull screeched. Beak open wide, he emitted a symphony of squawks, challenging the attacker. Undeterred, the second seagull hesitated, torn between his desire for snacks, and fear of the defending seagull. The tension hung in the air, and as the first seagull prepared to yell again, the second replied with a sullen "Fine," and flew off. Triumphant, the first gull tucked into his dinner.

Example 2. Detailed Description (Image)

**Instruction Prompt**

Explain all the details of the scene, starting from left to right.

Response

This is a photo of a small delivery truck. In the background of the photo is a loading dock of some type with green and yellow structures under an awning. The small truck stands at the center of the photo. It has a flatbed and awning over the flatbed and some type of white cabinet on the back and a black storage compartment under. The bed is green and has black and green milk crates stacked on it. The cab of the truck is small and tall with large windows all around. It is white with a green lower section and fine yellow stripe separating the green and white paint. The front of the truck has 'Dairy Crest' painted on it in green and has a European style license plate. There is a partial image of a man in the background.

Example 3. Open-ended Question and Answering (Video)

**Instruction Prompt**

What is this animal doing and why are they doing it?

Response

Pheasants eat gravel to help them in the digestion of their food. This helps them survive multiple weeks without feeding during extreme weather. A newly hatched chick will survive on insects for a few weeks of its life.

Example 4. Multimodal Reasoning (Audio)

Audio File (Freesound Link)

Instruction Prompt

What could be the likely momentous occasion in this sound clip?

Response

This is a countdown to a rocket taking off into space by a robotic voice.

Table 1: Example data from our MM-IT dataset, manually collected instruction prompt and response pairs for diverse modalities (*i.e.* image, video, audio). The collected instructions cover diverse domains (*e.g.* creative writing, open-ended reasoning), and are strictly grounded to the provided multimodal context (*i.e.* queries *cannot* be answered without understanding the accompanying multimodal context). The MM-IT data **serves as both a fine-tuning dataset as well as an evaluation benchmark** for complex multimodal reasoning tasks.

3.2 Fine-tuning with Multimodal Instruction Datasets

To further improve the model’s instruction-following capability with respect to diverse input modalities, we perform additional fine-tuning with our multimodal instruction-tuning (MM-IT) dataset. Specifically, we concatenate the input as [`<instruction> <modality_tokens>`], such that the response target is grounded on both textual instructions and the modality input. We perform ablations over (1) training the projection layers without altering the LLM parameters, or (2) using Low-Rank Adaptation [41] to further tune the LM behaviors.

We use both manually collected instruction-tuning datasets and synthetic data.

Manual Annotation. While there are publicly available third-party datasets on various VQA tasks, we observe that many of these data have insufficient diversity and quality — in particular for aligning LLMs towards diverse multimodal instruction-following tasks that go much beyond simple QA queries (*e.g.* “*Create a poem using this image*”, “*Extract the phone number on this flyer*”).

Therefore, we focus on collecting 60K examples of high-quality multimodal instruction tuning data for multiple modalities, as illustrated in Table 1. Specifically, we use various Creative Commons licensed, publicly available images, and augment these images with manually created instructions and responses. Annotators are required to provide instruction and answer pairs that are strictly multimodal, such that queries cannot be answered without understanding the accompanying multimodal context.

We show that our results notably improve using these fewer but well-balanced and higher-quality examples from our own vendor-based annotation efforts.

Synthetic Augmentation. In addition to the high-quality ground-truth instruction tuning data above, we augment the dataset using the LLaMA-2 (70B) [6] model, following similar approaches proposed by LLaVA [20]. Specifically, we use a textual representation of the image (*i.e.* multiple captions, bounding boxes information and objects) to generate question-answer pairs for the image. We generate 150K image-instruction-response pairs on varying domains and question types.

Note that our process strictly uses only open-sourced models – as opposed to other works that use commercial services such as ChatGPT or GPT-4.

4 Experiments

4.1 Tasks

We evaluate the model’s performance on two categories of tasks in the zero-shot setting: (1) captioning tasks for various modalities, and (2) multimodal reasoning and instruction-following tasks.

Captioning Tasks. We evaluate AnyMAL’s primary capability of generating captions given input modalities, which is aligned with the pre-training objective. The main purpose of the captioning task is to understand the alignment level between the text and other modalities after pre-training. Since the captioning tasks typically don’t require secondary reasoning steps, we expect that LLM weights or parameter sizes have less influence on the task.

Multimodal Reasoning Tasks. Given the high-level of alignment among the modalities, we evaluate the model’s reasoning and instruction-following abilities which it inherits from the core instruction-tuned LLM, as well as from the multimodal instruction-tuning process.

We conduct a comprehensive comparison with strong baseline models for each respective modality pair (vision-language and audio-language) from the open-sourced literature.

Note: As the MM-IT datasets include some in-domain images from public benchmarks (*e.g.* COCO), we report results separately for the pre-trained models (without further instruction tuning in Section 3.2) and the instruction-tuned models – to denote a *strict* zeroshot setup. All multimodal-instruction-tuned AnyMAL models are marked with “MM-IT” in the following sections.

4.2 Quantitative Analysis

Image Caption Generation: Table 2 shows zeroshot image captioning performance on COCO [48] and a subset of the MM-IT dataset marked with the “detailed description” task (MM-IT-Cap). It can be seen that our AnyMAL variants significantly outperform the baselines in both datasets. It is worthwhile to note that there is no significant gap between the performance of the AnyMAL-13B and the AnyMAL-70B variants. This result indicates that the underlying LLM capability has smaller impact to the image caption generation task (which corresponds to the core visual understanding capability), but is largely dependent on the scale of the data and the alignment methods. We attribute

Models	COCO	MM-IT-Cap
	CIDEr	CIDEr
BLIP-2	[3]	-
MiniGPT4	[21]	-
LLaVA	[20]	-
CM3Leon	[42]	61.6
OpenFlamingo-v2 9B	[17]	79.5
Flamingo-3B	[2]	73.0
Flamingo-9B	[2]	79.4
Flamingo-80B	[2]	84.3
IDEFICS-9B	[5]	46.0
IDEFICS-80B	[5]	91.8
AnyMAL 13B (ViT-G)	99.5	<u>15.5</u>
AnyMAL 70B (ViT-G)	<u>95.9</u>	15.7

Table 2: **Zeroshot Image Captioning** performance on COCO and MM-IT-Cap. Ablations (bottom) over our AnyMAL with varying LLM sizes. **Bold** and underlined denote the top and the second-best performance, respectively. “-”: the model (a) does not report results on the marked benchmarks, or (b) is pretrained or fine-tuned on the respective dataset, thus not suitable for the zeroshot evaluation above. AnyMAL **demonstrates the state-of-the-art zeroshot visual understanding capabilities compared to the baseline vision-language models.**

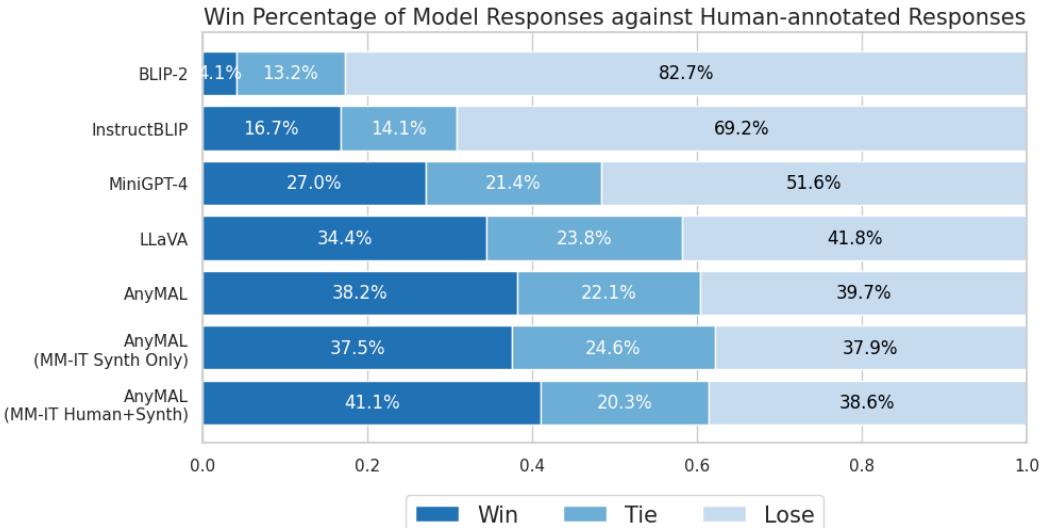


Figure 3: **Image-based reasoning** human evaluation results on pairwise comparisons (% win, tie and lose) with baseline outputs *against* the manually annotated ground-truth samples from MM-IT (1K test set). Baselines used: BLIP-2 (FlanT5_{XXL}) [3], InstructBLIP (Vicuna-13B) [19], MiniGPT4 [21] and LLaVA [20]. AnyMAL **demonstrates a smaller gap with human-generated responses** (41.1% win), compared to the baselines (LLaVA: 34.4% win, and MiniGPT4: 27.0%).

the slight under-performance of the AnyMAL-70B on COCO to the general verbosity of the LLaMA-70B model, which negatively impacts the score when evaluated against COCO captions that tend to be brief and concise. As expected, the automatic evaluation on MM-IT-Cap shows lower CIDEr scores overall, attributed to the much longer response length in detailed descriptions (See Table 1 for an example).

Human Evaluation on Multimodal Reasoning Tasks: MM-IT features diverse multimodal instruction and ground-truth answer pairs. We evaluate the performance of our models (pre-trained and instruction-tuned) against other vision-language models publicly available to run and use (*i.e.* LLaVA [20], MiniGPT4 [21]). Since the responses are subjective in nature (*e.g.* creative writing – “Write a

Models		Response Acc	Obj Recognition	Integrity
BLIP-2 (FlanT5 _{XXL})	[3]	43.3	73.5	99.3
InstructBLIP (Vicuna-13B)	[19]	46.3	73.2	98.3
Mini-GPT4	[21]	42.7	73.0	<u>99.5</u>
LLaVA	[20]	51.7	85.4	<u>99.5</u>
AnyMAL 70B		<u>56.0</u>	82.4	99.3
AnyMAL 70B (MM-IT Synth Only)		54.2	<u>83.5</u>	<u>99.5</u>
AnyMAL 70B (MM-IT Human+Synth)		58.0	79.3	99.7

Table 3: **Image-based Reasoning** human evaluation results on 1K test set from MM-IT on different axes: (a) Response Accuracy and Relevance (%) – whether responses are relevant to instructions and factually correct without any hallucinations, (b) Object Recognition (%) – whether key objects are identified at a detailed level, and (c) Integrity (%) – whether responses include offensive language. MM-IT indicates the model that has been instruction-tuned either with synthetic data only, or with the manually collected set (Section 3.2).

Models	H-Meme	VQAv2	TextVQA	S-QA	VizWiz	OKVQA
	AUC	Accuracy	Accuracy	Accuracy	Accuracy	Accuracy
OpenFlamingo-v2	[17]	51.6	50.5	24.2	-	27.5
Flamingo-3B	[2]	53.7	49.2	30.1	-	28.9
Flamingo-9B	[2]	57.0	51.8	31.8	-	28.8
Flamingo-80B	[2]	46.4	56.3	<u>35.0</u>	-	31.6
BLIP-2 (FlanT5 _{XXL})	[3]	52.0	65.0 [†]	44.1*	64.5	29.4
InstructBLIP (V-13B)	[19]	53.7	-	50.7 ^{†*}	<u>70.6</u>	33.4
IBELICS-9B	[5]	51.8	50.9	25.9	-	35.5
IBELICS-80B	[5]	60.6	<u>60.0</u>	30.9	-	<u>36.0</u>
AnyMAL 13B (ViT-G)		66.0	59.6	24.7	52.7	24.4
AnyMAL 70B (DINO-V2)		65.6	59.2	13.7	64.7	23.6
AnyMAL 70B (ViT-L)		<u>68.2</u>	62.0	35.4	67.2	32.2
AnyMAL 70B (ViT-G)		69.1	64.2	32.9	70.8	33.8
AnyMAL 70B (MM-IT;ViT-G)		67.4	67.8 [†]	32.5	67.6	41.3
						46.1

Table 4: **Zeroshot Image-based QA** results on 6 different VQA datasets (H-Meme: Hateful Meme, S-QA: Science QA). Ablations (bottom) over AnyMAL with varying base ViTs and LLM sizes. MM-IT (last row) denotes the model fine-tuned on our instruction dataset. **Bold** and underlined denote the top and the second-best performance, respectively. AnyMAL demonstrates competitive zeroshot multimodal reasoning capabilities, compared to the baseline vision-language models. *: Results with additional OCR inputs. †: in-domain images (*i.e.* COCO, TextCap) have been used during training, thus not a strict zeroshot performance.

poem about this image”, we believe that **human assessment** provides the most precise insight into the performance and capabilities of our proposed model.

We therefore collect pairwise comparisons for each baseline against 1K ground-truth samples (Figure 3), as well as the Likert scale scores (0-2) for each of the following criteria. The criteria for preference ranking includes response accuracy, object recognition accuracy, and integrity (see the full rubrics in Appendix A). Response accuracy measures whether the response contains the relevant, factually correct and verifiable information (without any hallucinations) with regards to the image and the instruction. Object recognition accuracy strictly measures whether the key objects are correctly recognized at a detailed level – primarily concerning the model’s visual knowledge. Finally, the integrity metric measures whether the response shows any harmful or offensive language.

Figure 3 shows that AnyMAL achieves strong performance with a narrower gap against the manually annotated ground-truth samples (41.1% win), compared to the baselines (LLaVA : 34.4% win, and MiniGPT4: 27.0% win). Notably, the model fine-tuned with the full instruction set exhibits the highest rate of preferential wins, showing a competitive level of visual understanding and reasoning capabilities **comparable to human-annotated responses**. It is also worthwhile to note that BLIP-2 and InstructBLIP suffer on these open-ended queries (4.1% and 16.7% preferential win, respectively), despite their strong performance in the public VQA benchmarks (Table 4).

Models	AudioCaps		
	CIDEr	SPICE	SPICEr
TopDown-AlignedAtt [36]	59.3	14.4	36.9
CNN10-VGG [43]	66.0	16.8	41.4
ACT [44]	67.9	16.0	42.0
PANNs + BERT [45]	66.7	17.2	42.0
AnyMAL 7B (CLAP)	70.4	21.0	45.7
AnyMAL 13B (CLAP)	<u>72.1</u>	<u>22.0</u>	<u>47.0</u>
AnyMAL 70B (CLAP)	77.8	23.0	50.4

Table 5: **Zeroshot Audio Captioning** results on AudioCaps. Ablations (bottom) over our AnyMAL with varying base LLMs and sizes. AnyMAL attains the best performance across multiple metrics, showing the model’s strong performance in audio signal understanding.

Models		STAR	How2QA	NextQA
		Accuracy	Accuracy	Accuracy
Internvideo (8)	[46]	41.6	62.2	49.1
Flamingo-9B	[2]	41.8	-	-
Flamingo-80B	[2]	39.7	-	-
BLIPv2 ViTG FlanT5xxl (4)	[47]	42.2	69.8	62.4
AnyMAL-Video 13B (Internvideo) (8)		37.5	54.8	46.8
AnyMAL-Video 70B (Internvideo) (8)		41.3	60	50.6
AnyMAL-Image 13B (ViT-G) (4)		<u>44.4</u>	59.6	47.9
AnyMAL-Image 70B (ViT-G) (4)		48.2	<u>68.1</u>	<u>57.6</u>

Table 6: **Zeroshot Video-based QA** accuracy on STAR, How2QA, and NextQA. Ablations (bottom) over AnyMAL with image vs video model and LLM sizes. AnyMAL demonstrates competitive zeroshot multimodal reasoning capabilities, compared to the baseline vision-language models. Numbers in brackets denote number of frames used per video clip.

Table 3 shows the detailed breakdown of scores for each criteria. Specifically, it can be seen that the model instruction-tuned with both manually collected and synthetically curated sets attains the highest response accuracy and relevancy score (12.2% relative improvement compared to the strongest baseline: LLaVA). This result highlights the enhanced capability of the model to comprehend and precisely answer questions in accordance with provided instructions. It is worth mentioning that the model, after instruction tuning, exhibits a decline in its ability to provide detailed recognition and descriptions of objects. We attribute this result to our intention behind collecting the AnyMAL dataset which promotes concise responses. Consequently, the model tends to provide more succinct descriptions, rather than elaborating on details (which often increases the risk of factual inaccuracies). High integrity scores (99+%) are achieved by all baselines.

VQA Benchmarks: Table 4 shows the zeroshot performance on the Hateful Meme dataset [49], VQAv2 [14], TextVQA [50], ScienceQA [51] (image-paired subset), VizWiz [52], and OKVQA [53] compared against the models in the literature that report zeroshot results on the respective benchmark. We focus on zeroshot evaluation to best estimate the model’s performance on the open-ended queries at inference time.

Overall, our AnyMAL exhibits the strongest performance compared to the baselines across multiple tasks. Notably, the pretrained AnyMAL models show competitive performance even without further fine-tuning on MM-IT – demonstrating strong reasoning capabilities retained in the base pre-training phase. Comparing the two text-guided vision encoders ViT-L [30] and ViT-G [31], ViT-G achieves higher scores in most of the vision datasets. These results suggest that larger encoders trained over more steps offer better representations. However, the TextVQA dataset presents a unique case in which the image encoder’s resolution plays a pivotal role in recognizing text within images. Consequently, the ViT-L model, with a resolution of 336x336, achieves a higher score than the ViT-G model, which had a resolution of 224x224. DinoV2 [54], which is trained in a discriminative self-supervised approach, shows worse performance compared to the text-guided vision encoders,

Audio Signals Input	AnyMAL Output
	<p>Gold Caption: Typing on a keyboard is occurring in a quiet environment</p> <p>Prompt: Can you guess my location? Reply with one or two sentences.</p> <p>Response: Based on the sound of the typing, I would guess that you are located in a quiet, indoor setting, possibly in an office or study. The sound of the typing is clear and crisp, with a slight echo or reverb that suggests a room with a high ceiling or a large, open space.</p>
	<p>Gold Caption: Very strong wind is blowing, and waves are rustling</p> <p>Prompt: Based on the weather, what should I wear? Reply with one or two sentences.</p> <p>Response: Based on the sound of the wind and the waves, it seems like a stormy weather is approaching. It's best to dress in layers and wear something waterproof and warm to stay comfortable and safe.</p>
	<p>Gold Caption: A baby cries while electronic speech occurs in the background, an adult female speaks, a sneeze occurs, and an adult female speaks once more.</p> <p>Prompt: Describe the sound in one or two sentences.</p> <p>Response: The sound is a loud, high-pitched scream of a child, with a hint of a whine, as if the child is crying and screaming at the same time. The sound is accompanied by the muffled voices of adults in the background, possibly the child's parents or caregivers, speaking in a soothing tone.</p>

Table 7: Example outputs from AnyMAL on unseen audio samples. Gold captions are shown as a reference for the audio signals. The text prompt and the audio signals are used as input.

demonstrating the importance of feature space alignment. Among the base LLM models, our 70B model demonstrates the most robust performance, underscoring the influence of substantial reasoning proficiency inherent in larger LLMs on tasks involving visual reasoning.

Video QA benchmarks: We evaluate our model on three challenging video question-answering benchmarks in Table 6: How2QA [55], STAR [56] and NextQA [57]. Our model demonstrates competitive results compared to the baselines, and achieves state-of-the-art performance on the STAR benchmark. Note that we compare against approaches that process the full, untrimmed video clip to generate answers. Prior work has shown additional improvements with careful frame-selection strategies [58]. Our approach is compatible with such strategies, however that is beyond the scope of our experiments. We additionally report model variants trained exclusively on videos from HowTo100M [59] coupled with text from ASR transcripts, and using explicit video encoders (Internvideo [46]) as opposed to image encoders. However, these models perform worse due to the weak alignment of ASR to video clips and lower diversity in content (i.e., instructional videos).

Ablations on Hyperparameters: Figure 4 shows the training losses on the variants of AnyMAL during image-text pre-training. Due to the high computational cost of training 70B models, we conduct ablations only on the 13B models. We then use the optimal set of hyperparameters from these experiments as the default configuration for

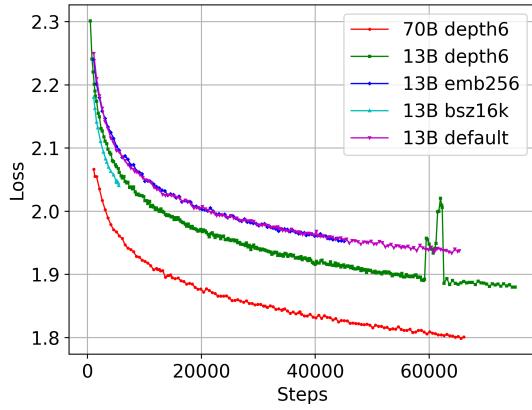


Figure 4: Training losses on variants (13B and 70B) of AnyMAL during image-text pre-training. Default parameters: batch size: 2,048, resampler layers: 2, the number of modality tokens: 64. Each line features an ablation with one differing parameter from the default setting, *i.e.* bsz16K: 16,384 batch size, emb256: 256 visual tokens, depth6: 6 resampler layers.

IMU Signals Input	Reference Video	AnyMAL Output
		Prompt: Describe the motion. Response: [Camera wearer] looks around.
		Prompt: Describe the motion. Response: [Camera wearer] walks in the house.
		Prompt: What's the safest way to stop? Response: To stop safely on a bike, use the brakes gradually and look where you want to go. Lean the bike slightly to one side and keep your body upright.

Table 8: Example outputs from AnyMAL, on IMU motion sensor samples from Ego4D. The text prompt and the IMU signals are used as input. We display a few frames from the synchronized video as a reference for readers. The reference video is not used as input to the model.

training the 70B models. Since the trainable parameters remain consistent for both the 13B and 70B models, we anticipate that the behavior during model training will remain unchanged. The following key hyperparameters are considered: the layers of the Resampler (2 vs. 6), the number of multimodal token embeddings to represent visual input (64 vs. 256 tokens), and batch size of pre-training (2,048 vs. 16,384). Overall, we observed that increasing the batch size and the number of visual tokens yields minimal improvement. In contrast, increasing the number of resampling layers significantly reduces the loss without substantially increasing the training budget.

Effects of Scaling LLM Parameter Size (70B vs. 13B): The 70B model demonstrates a reduced training loss overall when compared with the 13B versions. This loss performance also aligns with the downstream task results in Tables 2 and 4. We attribute this result to the inherent reasoning abilities and the knowledge assimilated within the 70B models, which expedites the visual concept acquisition and alignment process. Overall, the comparison demonstrates the importance of scaling LLM parameters in vision-language pre-training as well, which is an aspect that has seldom been addressed in existing literature.

Audio Caption Generation: Table 5 shows the audio captioning results on the AudioCaps [36] benchmark dataset. AnyMAL significantly outperforms other state-of-the-art audio captioning models in the literature (*e.g.* +10.9pp in CIDEr, +5.8pp in SPICE), showing the versatility of the proposed approach on various modalities beyond just vision. We note that our 70B model displays notably strong performance compared to the 7B and the 13B variants – showing the importance of the reasoning module for the task.

IMU Motion Description Generation: We use the Ego4D [38] dataset to train an IMU-aligned AnyMAL-7B model, leveraging the synchronized IMU sensor data and textual narrations provided in the dataset. Given that the task of generating textual descriptions from motion signals has not been previously achievable or reported, we solely present the performance achieved by our own model.

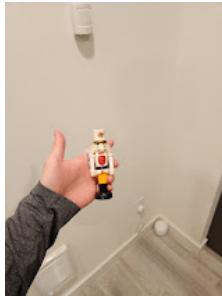
On the held-out test set, we achieve 52.5 CIDEr and 23.2 ROUGE-L against the ground-truth captions, showing the feasibility of the newly proposed task.

Combining this captioning ability with the reasoning capabilities of LLMs, in Table 8 we show examples of *novel* applications that AnyMAL might allow, *e.g.* inferring user motion states and incorporating these as part of its response (*e.g.* “What’s the safest way to stop?” → “To stop safely on a bike, ...” without any textual or visual cues that the user is biking).

4.3 Qualitative Analysis

Comparison with other Vision-Language Models: Table 9 and 10 show outputs from various vision-language models [17, 19, 20, 21] on diverse example image and prompt pairs, compared with AnyMAL (LLaVA-70B). It can be seen that AnyMAL exhibits strong visual understanding capabilities (such as identification of objects and their states), as well as language generation capabilities. While MiniGPT4 [21] and LLaVA [20] do present reasonable and fluent responses, their accuracy is not

Example 1. Creative Writing

Input Image

Prompt	<i>Write a joke about it.</i>
BLIP-2	a person holding a lego toy in their hand
InstructBLIP	a hand holding a small toy in a room
Mini-GPT4	Why did the lego man break up with his lego wife? Because he found a better lego man.
OpenFlamingo	What do you get when you cross a snowman with a robot? Answer: A snowbot.
LLaVA	A British solider mannequin has lost its way in a New York City apartment. When the surprised tenant, who doesn't seem to know much about British (...)
AnyMAL 70B	Here's a joke for you: Why did the nutcracker doll go to the doctor? Because he was feeling a little cracked!

Example 2. How-To

Input Image

Prompt	<i>How do I fix it?</i>
BLIP-2	a bike that has been thrown off a cliff and landed on a rock in the middle of a field
InstructBLIP	tire
Mini-GPT4	The bike has a flat tire. To fix it, you will need to pump up the tire with air. First, remove the wheel from the bike and place it on the ground. Then, (...)
OpenFlamingo	This is a flat tire.
LLaVA	It appears that the person biking along the way has experienced a flat tire because their bike is leaning against a wall, and there is a bike tube lying on the ground. To fix the issue, the person would need to first examine the tire to determine the location of the flat. They should then remove the damaged tube from the tire. (...)
AnyMAL 70B	It appears that the front tire of one bicycle is flat. You might want to try changing the inner tube or patching any holes in the existing one before re-inflating the tire.

Table 9: Examples comparing AnyMAL with various baselines: BLIP-2 (FlanT5_XXL) [3], InstructBLIP (Vicuna-13B) [19], OpenFlamingo-v2-9B [17], Mini-GPT4 [21], and LLaVA [20]. AnyMAL demonstrates competitive performance in object recognition (*e.g.* “nutcracker”), response relevancy (*e.g.* writing a joke, explaining how-to), and secondary reasoning.

Example 3. Recommendation

Input Image


Prompt	<i>Which one would pair better with steak?</i>
BLIP-2	red wine or white wine?
InstructBLIP	red 1
Mini-GPT4	The wine bottles in the image are two different types of wine, one is a red wine and the other is a white wine. Both of these wines would pair well with steak. The red wine would be a good choice for a rich and flavorful steak, while the white wine would be a good choice for a lighter steak or seafood dish.
LLaVA	For a steak, the red wine is likely a more appropriate pairing compared to the white wine. Whether it is Chianti or Chianti Classico, both options are Italian red wines made from Sangiovese grapes, and their full-bodied nature provides a rich flavor profile, making them good companions to steak dishes. The white wine, Soave Classico DOC, is a white grape varietal, which is less traditional for pairing (...)
AnyMAL 70B	The Barolo on the left would pair better with steak. It's a full-bodied, high tannin wine that can stand up to the richness of red meat.

Example 4. Question and Answering

Input Image


Prompt	<i>How long does this river go?</i>
BLIP-2	it's a river that runs through the city of florence, tuscany, italy
InstructBLIP	10 miles
Mini-GPT4	The river goes over the bridge in the image.
LLaVA	The river in the image is quite long, stretching across the city and passing under several buildings, including a bridge.
AnyMAL 70B	The river in the image is the Arno River, which flows through Florence, Italy. It stretches for approximately 241 kilometers (150 miles) and empties into the Ligurian Sea.

Table 10: Examples comparing AnyMAL with various baselines. AnyMAL demonstrates competitive performance in both object recognition (*e.g.* “Barolo”, “Arno River”) and secondary reasoning.

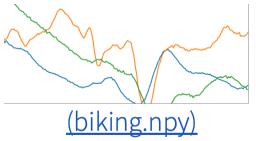
Input Modality 1	Input Modality 2	AnyMAL Output
 <u>(biking.npy)</u>		<p>Prompt: (Given the motion signals) Write a social media caption for this view.</p> <p>Response: Pedaling along the San Francisco Bay, taking in breathtaking views of the <u>Golden Gate Bridge</u>! The salty sea breeze <u>invigorates me as I make my way to my next destination</u>.</p>

Table 11: Example outputs from AnyMAL, with multiple **interleaved modalities** as input. The text prompt and two other modalities (*e.g.* image & IMU motion sensor signals) are used as input. The underlined text in the response demonstrates the evidence that the output is grounded on multiple modalities.

consistently ensured. These examples effectively highlight the benefits of the proposed approach which allows for large-scale pre-training covering diverse visual concepts, while inheriting strong reasoning capabilities derived from instruction-tuned LLMs.

We note that we use the latest checkpoints made available for each baseline to generate responses.

Interleaved Modalities: The flexible model architecture of AnyMAL allows for combinatory modalities as conditioning context (*e.g.* image + IMU motion sensor signals), which allows for more comprehensive multimodal reasoning. We demonstrate the model’s zeroshot capabilities of handling such interleaved modalities in Table 11 (*e.g.* composing a message with a given image (Golden Gate Bridge), with the user’s prevalent motion (biking) as part of the context).

This result illustrates the new and natural way of interaction with an AI model made possible by AnyMAL, wherein a user can presume a shared understanding of combined sensory perceptions (*e.g.* visual, auditory, and motion cues) when composing queries – avoiding the need to specify multimodal contexts.

5 Safety

Inference Time Integrity. To ensure the safety and integrity of the AnyMAL model, several measures are made on the following categories of potential integrity violations: (1) input images, (2) input text prompts, (3) text outputs, and (4) multimodal combination of input images and text outputs.

- (1) Input image: we use a pre-trained image classifier based on RegNetY [60] to detect any content that violates integrity standards. This detection encompasses graphic material, violent imagery, hate symbols, instances of bullying, harassment, *etc.* If such a violation is identified within the image, we proceed to reject the entire query.
- (2) Input text prompt: we use a RoBERTa-based text classifier [61] trained to detect integrity-violating utterances such as violence, harassment, hate speech, *etc.* When a violation is detected in user prompt, we proceed to reject the entire query.
- (3) Output text: we employ the same text classifier in (b) to detect any problems within the generated output. For streaming use cases, we run the classifier for each sentence to promptly identify any violations.
- (4) Multimodal association in input image & output text: in the uncommon scenario where harmless text and a harmless image (which individually appear innocuous) can result in a problem when they are associated, we use a multimodal classifier to detect such instances.

Training Time Safety. The datasets used for pre-training (*e.g.* [34, 62]) have gone through a filtration process to remove harmful language or images that compromise integrity, thereby reducing the potential for the model to generate content that violates integrity standards.

LLM Safety. Since our AnyMAL pre-training does not alter the parameters of the base LLM, we carry over the same safety precautions implemented for its language generation. For instance, LLaMA-2 (the version we report most of our results on) places safeguards such as negative example fine-tuning, reinforcement learning with human feedback (RLHF) [63, 64, 65].

6 Conclusions

Our proposed AnyMAL showcases a novel and natural way of interacting with an AI model, *e.g.* asking questions that presume a shared understanding of the world between the user and the agent, through the same lens and combinatory perceptions (*e.g.* visual, auditory, and motion cues). The proposed scalable way of training AnyMAL makes it possible to leverage the powerful reasoning capabilities of the LLaMA-2 language model within the multimodal settings.

Our contributions are as follows: (1) We present a large-scale Multimodal LLM (AnyMAL), trained using open-sourced resources and scalable solutions for multiple modalities. (2) We introduce the Multimodal Instruction Tuning dataset (MM-IT), a first-of-its-kind collection of high-quality manual annotations of multimodal instruction data. (3) Our comprehensive empirical analysis shows insights to the efficient and scalable *recipe* for building a multimodal reasoning model, given various LLMs and modeling choices.

7 Limitations

We discuss the current limitations of our work as follows.

First, the proposed causal multimodal language modeling approach still encounters challenges in establishing a robust grounding with the input modality. Specifically, we observe that during the generation, the model occasionally prioritizes focusing more on the generated text rather than the input image. This leads to the generation of output that incorporates biases acquired from the underlying language model (LLM), which can incur inaccuracies when compared against the image context. We expect that additional architectural adjustments or unfreezing LLM parameters are necessary to address this limitation effectively (albeit the much higher computational costs it might entail).

Second, while we greatly increase the size of the pretraining dataset, the understanding of visual concepts and entities remains constrained by the quantity of paired image-text data included in the training process. In the domain of text-only language models, it is commonly observed that approaches incorporating external knowledge retrieval significantly enhance the model’s ability to overcome its knowledge limitations. These approaches offer a potential means to alleviate the limitations mentioned earlier.

Lastly, in the scope of our work, the multimodal adaptation of an LLM is bounded by four modalities: image, video, audio, and IMU signals. While we believe that the proposed approach has the potential to encompass any other modality, provided there exists a paired dataset, its effectiveness for such modalities still needs to be substantiated.

References

- [1] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, “Multimodal few-shot learning with frozen language models,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.
- [2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, “Flamingo: a visual language model for few-shot learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
- [3] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.
- [4] OpenAI, “Gpt-4 technical report,” *ArXiv*, vol. abs/2303.08774, 2023.
- [5] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh, “Obelics: An open web-scale filtered dataset of interleaved image-text documents,” 2023.
- [6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, “Language models are unsupervised multitask learners,” 2019.

- [8] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, “Scaling instruction-finetuned language models,” 2022.
- [9] B. Wang and A. Komatsuzaki, “GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.” <https://github.com/kingoflolz/mesh-transformer-jax>, May 2021.
- [10] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, “Opt: Open pre-trained transformer language models,” *arXiv preprint arXiv:2205.01068*, 2022.
- [11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [12] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” March 2023.
- [13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in *International conference on machine learning*, pp. 2048–2057, PMLR, 2015.
- [14] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, “VQA: Visual question answering,” in *ICCV*, 2015.
- [15] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, “Visual dialog,” in *CVPR*, 2017.
- [16] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2018.
- [17] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, *et al.*, “Openflamingo: An open-source framework for training large autoregressive vision-language models,” *arXiv preprint arXiv:2308.01390*, 2023.
- [18] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. Florence, “Palm-e: An embodied multimodal language model,” in *arXiv preprint arXiv:2303.03378*, 2023.
- [19] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” 2023.
- [20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” 2023.
- [21] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.
- [22] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, “Otter: A multi-modal model with in-context instruction tuning,” *arXiv preprint arXiv:2305.03726*, 2023.
- [23] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, *et al.*, “mplug-owl: Modularization empowers large language models with multimodality,” *arXiv preprint arXiv:2304.14178*, 2023.
- [24] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, “Multimodal-gpt: A vision and language model for dialogue with humans,” *arXiv preprint arXiv:2305.04790*, 2023.

- [25] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, H. Li, and Y. Qiao, “Llama-adapter v2: Parameter-efficient visual instruction model,” *arXiv preprint arXiv:2304.15010*, 2023.
- [26] H. Zhang, X. Li, and L. Bing, “Video-llama: An instruction-tuned audio-visual language model for video understanding,” *arXiv preprint arXiv:2306.02858*, 2023.
- [27] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, “Pandagpt: One model to instruction-follow them all,” *arXiv preprint arXiv:2305.16355*, 2023.
- [28] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, “Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration,” *arXiv preprint arXiv:2306.09093*, 2023.
- [29] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, *et al.*, “Mit: A large-scale dataset towards multi-modal multilingual instruction tuning,” *arXiv preprint arXiv:2306.04387*, 2023.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, “Learning transferable visual models from natural language supervision,” in *International Conference on Machine Learning (ICML)*, 2021.
- [31] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, “Laion-5b: An open large-scale dataset for training next generation image-text models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.
- [32] Y. Wu*, K. Chen*, T. Zhang*, Y. Hui*, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [33] S. Moon, A. Madotto, Z. Lin, A. Dirafzoon, A. Saraf, A. Bearman, and B. Damavandi, “Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text,” *arXiv preprint arXiv:2210.14395*, 2022.
- [34] F. Radenovic, A. Dubey, A. Kadian, T. Mihaylov, S. Vandenhende, Y. Patel, Y. Wen, V. Ramanathan, and D. Mahajan, “Filtering, distillation, and hard negatives for vision-language pre-training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6967–6977, 2023.
- [35] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *Proc. IEEE ICASSP 2017*, (New Orleans, LA), 2017.
- [36] C. D. Kim, B. Kim, H. Lee, and G. Kim, “Audioscaps: 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)*, pp. 119–132, 2019.
- [37] K. Drossos, S. Lipping, and T. Virtanen, “Clotho: An audio captioning dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, IEEE, 2020.
- [38] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erappalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik, “Ego4d: Around the World in 3,000 Hours of Egocentric Video,” in *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

- [39] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, A. Desmaison, C. Balioglu, P. Damania, B. Nguyen, G. Chauhan, Y. Hao, A. Mathews, and S. Li, “Pytorch fsdp: Experiences on scaling fully sharded data parallel,” 2023.
- [40] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *arXiv preprint arXiv:2305.14314*, 2023.
- [41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [42] L. Yu, B. Shi, R. Pasunuru, B. Miller, O. Golovneva, T. Wang, A. Babu, B. Tang, B. Karrer, S. Sheynin, C. Ross, A. Polyak, R. Howes, V. Sharma, J. Xu, U. Singer, D. Li, G. Ghosh, Y. Taigman, M. Fazel-Zarandi, A. Celikyilmaz, L. Zettlemoyer, and A. Aghajanyan, “Scaling autoregressive multi-modal models: Pretraining and instruction tuning,” 2023.
- [43] X. Xu, H. Dinkel, M. Wu, Z. Xie, and K. Yu, “Investigating local and global information for automated audio captioning with transfer learning,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 905–909, IEEE, 2021.
- [44] X. Mei, X. Liu, Q. Huang, M. D. Plumbley, and W. Wang, “Audio captioning transformer,” *arXiv preprint arXiv:2107.09817*, 2021.
- [45] X. Liu, X. Mei, Q. Huang, J. Sun, J. Zhao, H. Liu, M. D. Plumbley, V. Kilic, and W. Wang, “Leveraging pre-trained bert for audio captioning,” in *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1145–1149, IEEE, 2022.
- [46] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, *et al.*, “Internvideo: General video foundation models via generative and discriminative learning,” *arXiv preprint arXiv:2212.03191*, 2022.
- [47] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.
- [48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *ECCV*, 2014.
- [49] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, “The hateful memes challenge: Detecting hate speech in multimodal memes,” *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.
- [50] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach, “Towards vqa models that can read,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.
- [51] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to explain: Multimodal reasoning via thought chains for science question answering,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.
- [52] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, “Vizwiz grand challenge: Answering visual questions from blind people,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.
- [53] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, “Ok-vqa: A visual question answering benchmark requiring external knowledge,” in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.
- [54] M. Oquab, T. Dariseti, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2023.
- [55] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, “Hero: Hierarchical encoder for video+language omni-representation pre-training,” *arXiv preprint arXiv:2005.00200*, 2020.

- [56] B. Wu, S. Yu, Z. Chen, J. B. Tenenbaum, and C. Gan, “Star: A benchmark for situated reasoning in real-world videos,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [57] J. Xiao, X. Shang, A. Yao, and T.-S. Chua, “Next-qa: Next phase of question-answering to explaining temporal actions,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.
- [58] S. Yu, J. Cho, P. Yadav, and M. Bansal, “Self-chained image-language model for video localization and question answering,” *arXiv preprint arXiv:2305.06988*, 2023.
- [59] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.
- [60] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, “Designing network design spaces,” 2020.
- [61] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [62] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, *et al.*, “Make-a-video: Text-to-video generation without text-video data,” *arXiv preprint arXiv:2209.14792*, 2022.
- [63] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” *Advances in neural information processing systems*, vol. 30, 2017.
- [64] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhosseini, C. McKinnon, *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [65] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *arXiv preprint arXiv:2305.18290*, 2023.
- [66] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Oct. 2020.
- [67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Neural Information Processing Systems*, 2019.

A Human Evaluation

To measure if the AnyMAL response is relevant and correct with respect to the query and the image, we use the following rubrics to collect human feedback data.

Recognition Accuracy: Did the model correctly understand and identify the objects or scenarios in the image relevant to the query?

- 0: The model completely misinterprets the image or fails to recognize key objects in the image relevant to the query.
- 1: The model partially misinterprets the image or recognizes some but not all key elements.
- 2: The model correctly interprets the image and identifies key elements.

Response Accuracy: Does the response provide factually correct information? Does it Hallucinate?

- 0: The response contains factually incorrect information or even partially incorrect information.
- 2: The response provides factually accurate information.

Integrity: Does the response show any harmful or offensive language or signs of unjust or prejudiced treatment related to characteristics such as race, age, gender, nationality, etc.?

- 0: The response shows some sign of bias or harmful/offensive language.
- 2: The response shows no signs of bias or harmful/offensive language.

Pairwise Comparison: choose the response that better answers the given question, and that is more contextually appropriate and accurate (win, tie, or lose).

We plan to explore further training the model with the human feedback data (e.g. RLHF).

B Additional Notes on Experiments

Different prompts were used to get the model output in the desired format for each task (e.g. multiple choice questions, yes/no questions). Below is the full list of prompts used for each task.

B.1 Multimodal Prompts

MM-IT System message: “*You are a multimodal assistant, designed to provide helpful answers to users’ image-related questions.* \n\n Here is the image: ”. User message: “{question}”

VQA, TextVQA, OKVQA System message: “*You are a multimodal assistant, designed to provide direct answers to users’ image-related questions. Reply directly with only one phrase. *Do not* start your answer with ‘Sure ...’.* \n\n Here is the image: ”. User message: “*In the image, {question} Reply in one word.*”

VizWiz System message: “*Answer the questions based on the image when possible, otherwise say ‘unanswerable’.* \n\n Here is the image: ”. User message: “*In the image, {question} Reply in one phrase/word or say ‘unanswerable’*”

Hateful Meme System message: “*You are a social media content moderator, designed to detect hateful memes.* \n\n Here is the meme: \n This meme contains text: ‘{ocrl}’”. User message: “*Is this a hateful meme? Answer yes or no.*”

Coco Caption System message: “*You are a multimodal assistant, designed to provide direct and concise answers to users’ image-related requests.* \n\n Here is the image: ”. User message: “*Describe the image with one *generic* sentence using json format. Here are two examples:*\n Specific: {"caption": "Body-Solid (Best Fitness) Inversion Table-2"}\n Generic: {"caption": "A man laying on top of an exercise table."}

ScienceQA System message: “*Given the image, choose the correct option for the following question. Your response must be just a single letter that corresponds to the correct option (e.g. A, B)* \n\n Here is the image: . User message: “{context} Question: {question}\n Options: {choices}\n\n Reply in a single letter.”

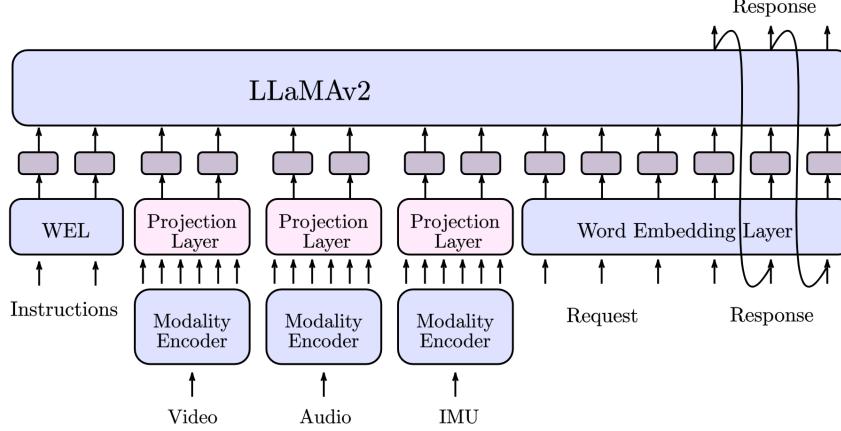


Figure 5: AnyMAL Inference example with multiple modality as input.

AudioCap System message: “*You are a multimodal assistant. Designed to provide direct answers to users’ audio-related questions. Here is the audio: <audio>*” User message: “*Describe the sound.*”

STAR, How2QA, NextQA System message: “*You are a multimodal assistant. Designed to provide direct answers to users’ video-related questions. \n\n Here is the video: <video>.*” User message: {question} Select exactly one option from the following: [options].

IMU-Ego4d System message: “*You are a multimodal assistant, designed to provide helpful, concise and direct answers to users’ questions, based on the user’s motion sensor signals reading from a head-mounted IMU device. The signals may indicate that a user may be running, walking, biking, driving, looking around, etc. Always answer under 30 words. \n\n Here are the user’s predicted motions: <IMU>*” User message: “*Describe this motion.*”

B.2 Multimodal Inputs

Figure 5 shows the diagram for performing modality-interleaved inference (for examples shown in Table 11).

B.3 Hyperparameters

Pre-training: Table 12 report the hyper-parameters used in this work for model pre-training.

Models	Batch Size	Initial LR	# Steps	# Modality Embeddings	Projection Module (#Layers)
AnyMAL (13B, Image)	2048	2×10^{-4}	100k	64	Resampler (6)
AnyMAL (70B, Image)	2048	2×10^{-4}	100k	64	Resampler (6)
AnyMAL (13B, Audio)	128	1×10^{-4}	1k	32	Linear (1)
AnyMAL (70B, Audio)	128	1×10^{-4}	1k	32	Linear (1)
AnyMAL (13B, Video)	1024	1×10^{-4}	20k	32	Resampler (4)
AnyMAL (70B, Video)	1024	1×10^{-4}	20k	32	Resampler (4)
AnyMAL (7B, IMU)	256	1×10^{-4}	2k	32	Linear (1)

Table 12: Hyperparameters for AnyMAL Pre-training

Finetuning: We finetune both resampler and LoRA adapters on MM-IT training set with the prompt described in B.1. Following the QLoRA [40] work, we set LoRA $r = 64$, $\alpha = 16$, and add LoRA modules on all linear layers. We finetune the model for 3k steps with batch size 128, initial LR 1×10^{-5} .

B.4 Code Base & Hardware

The implementations of the transformer-based models are extended from the HuggingFace² code base [66] and other cited authors’ released code-bases. Our entire code-base is implemented in PyTorch [67]. All models in this work are trained on a varying number of Nvidia A100³ GPUs.

B.5 More Qualitative Examples

We provide more qualitative examples in Table 13 and 14, comparing outputs from various baselines. While some of the baselines do present reasonable and fluent responses, their accuracy is not consistently ensured, either in their visual understanding (*e.g.* what objects are present in an image) or secondary reasoning (*e.g.* analyzing whether there is any element in an image that violates integrity policy).

²<https://github.com/huggingface/transformers>

³<https://www.nvidia.com/en-us/data-center/a100/>

Example 5. Recommendation - Fashion

Input Image

Prompt	<i>What jacket would go well with this?</i>
BLIP-2	blue shirt, black pants, grey shoes
InstructBLIP	black
Mini-GPT4	A light blue shirt would go well with this outfit.
OpenFlamingo	This jacket would go well with this outfit.
LLaVA	With the man wearing a blue shirt and a pair of black pants, a solid-colored blazer, such as a navy or black one, would go well with the outfit.
AnyMAL 70B	A navy blazer would look great with this shirt. The navy color will complement the shirt's blue hue, and the blazer will add a more formal touch to the outfit.

Example 6. Multimodal Integrity Violation Detection

Input Image

Prompt	<i>Is this meme hateful?</i>
BLIP-2	no
InstructBLIP	no
Mini-GPT4	No, it is not hateful. It is a joke that is meant to be humorous and not offensive.
OpenFlamingo	No, this meme is not hateful.
LLaVA	No, the meme is not hateful. It is a humorous and lighthearted image that features a toilet with a sign that reads "Hey Friend, Find a Perfect Home for You." The meme is intended to be amusing and relatable, as it highlights the importance of finding a comfortable and suitable living space. The image is not intended to offend or belittle anyone, but rather to bring a smile to the viewer's face.
AnyMAL 70B	The meme may offend the recipient by associating a toilet in a bathroom as a perfect home that the recipient should live in.

Table 13: More examples comparing AnyMAL with various baselines: BLIP-2 (FlanT5_XXL) [3], InstructBLIP (Vicuna-13B) [19], OpenFlamingo-v2-9B [17], Mini-GPT4 [21], and LLaVA [20].

Example 7. Inspiration - Recipe

Input Image

Prompt	<i>What cocktails can I make with these ingredients?</i>
BLIP-2	gin margarita gin margarita gin margarita
InstructBLIP	There are several cocktails that can be made with the ingredients in the image, including a classic Margarita, a Spritz, and a Gimlet. The Margarita is a classic cocktail made with tequila, lime juice, and triple sec, while the Spritz is a refreshing drink made with prosecco, Aperol, and soda water. (...)
Mini-GPT4	There are many cocktails that can be made with these ingredients. Here are a few examples:
OpenFlamingo	You can make a margarita, a cosmopolitan, and a whiskey sour.
LLaVA	With the items on the counter, you can make several cocktails. Some of the ingredients include a bottle of vodka, orange, apple, champagne, cognac, and wine. These can be used to create various cocktails like the following: (...)
AnyMAL 70B	Here are some ideas: 1) A classic vodka sour. 2) An orange blossom martini, with vodka, triple sec, and freshly-squeezed orange juice.

Example 8. Detailed Description

Input Image

Prompt	<i>Describe this image in detail.</i>
BLIP-2	a group of people sit around a campfire in the woods
InstructBLIP	The image depicts a group of people gathered around a campfire in a forest. There are six people sitting in chairs, with some standing nearby. There are several chairs placed in a semi-circle formation, with some people sitting on them and others standing nearby.
Mini-GPT4	This image shows a group of people sitting around a campfire in a forest. They are all wearing warm clothing and hats, and some of them are holding cups of hot drinks. The forest is filled with tall trees and the sky is clear and blue. The ground is covered in leaves and pine needles.
OpenFlamingo	This image shows a group of friends sitting around a campfire.
LLaVA	In this image, nine people, including three young children, aged 4, 5, and 11, are sitting in various chairs around an outdoor campfire, enjoying each other's company.
AnyMAL 70B	Four people are gathered around a campfire in the woods. They're all wearing casual outdoor clothing and sitting on chairs. There's a tent pitched nearby, and various camping gear scattered about.

Table 14: More examples comparing AnyMAL with various baselines: BLIP-2 (FlanT5_XXL) [3], InstructBLIP (Vicuna-13B) [19], OpenFlamingo-v2-9B [17], Mini-GPT4 [21], and LLaVA [20].