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-3 (70B), and converts modality-specific signals to the joint textual space through a pre-trained aligner module.

In this paper, we provide details on the optimizations implemented to efficiently scale the training pipeline, and present a comprehensive *recipe* for model and training configurations. We conduct comprehensive empirical analysis comprising both human and automatic evaluations, and demonstrate state-of-the-art performance on various multimodal tasks compared to industry-leading models – albeit with a relatively small number of trainable parameters.

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 (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Li et al., 2023b; OpenAI, 2023), 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 (Li et al., 2023b; Laurençon et al., 2023), such as text and image models, or has centered on proprietary language models that are not open sourced (Alayrac et al., 2022; OpenAI, 2023).

To tackle the previously mentioned challenges, we introduce **Any-Modality Augmented Language Model** (AnyMAL) — a collection of multimodal 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 (Tsimpoukelli et al., 2021) to (1) more capable instruction-tuned LLMs (*i.e.* Llama-3-70B-chat (AI@Meta, 2024)), (2) larger pre-trained modality encoders, and (3) advanced projection layers to handle variable input lengths. The output examples are shown in Figure 1, and an illustration of the methodology is in Figure 2.

The key contributions of the work are as follows:

- We present an efficient and scalable *recipe* for building Multimodal LLMs. We provide projection layers pre-trained on large datasets with diverse modalities (*e.g.* $500M$ images, $2.2M$ audio, $500K$ IMU time-series, $28M$ videos) all aligned to the same LLM (Llama-3-70B-chat), thus enabling interleaved multimodal in-context prompting.
- We fine-tune the model with the multimodal instruction set and human preference data 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.
- We provide details on the GPU optimization strategies implemented to scale the training pipeline to 70B models, and the recipe for model and training configurations.
- Our best model achieves strong zero-shot performance in both automatic and human evaluation on diverse tasks and modalities, setting new SOTA on MMBench, AI2D and MathVista, and +14.5% relative CIDEr improve-

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

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1314–1332
November 12-16, 2024 ©2024 Association for Computational Linguistics

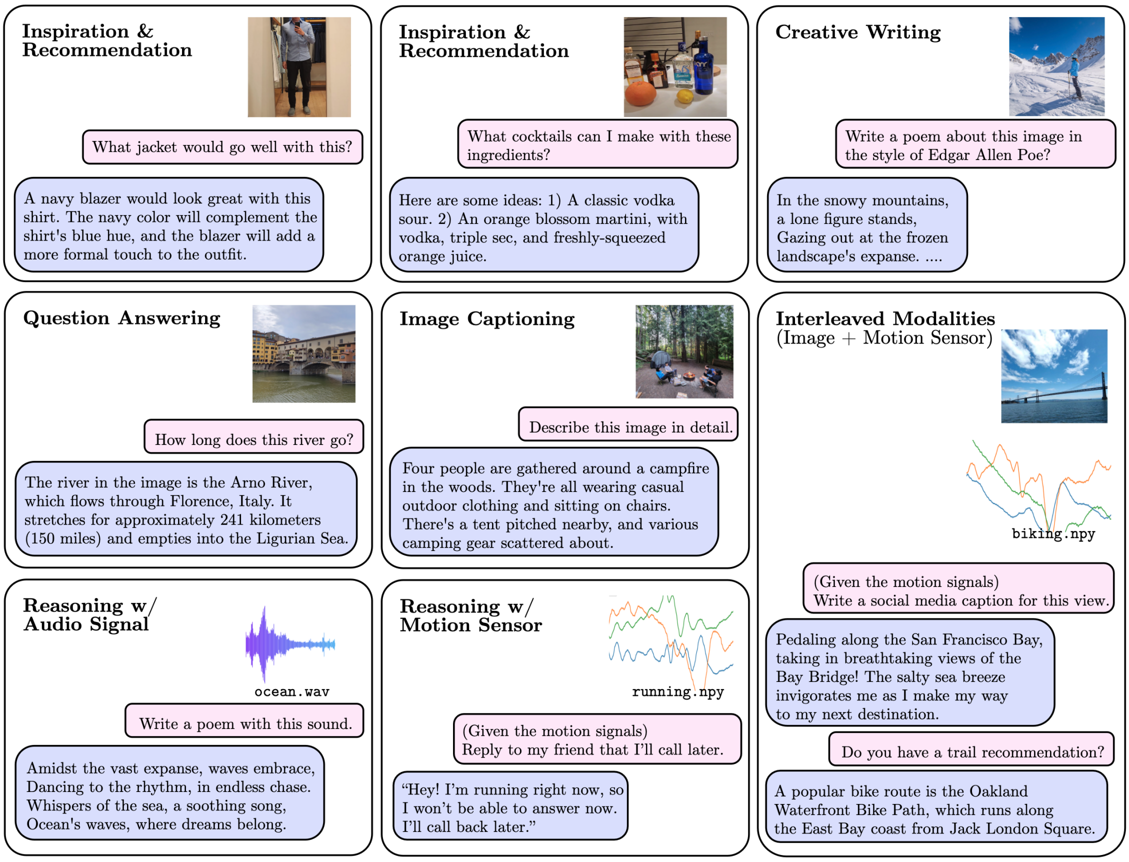


Figure 1: Example AnyMAL outputs. The model understands various input signals (*i.e.* vision, audio, motion sensor signals), and responds to free-form user queries. When multiple modalities are interleaved and given as input (*e.g.* right-most: image + IMU motion sensor signals), the model reasons over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

ment 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 GPT4 (OpenAI, 2023), the publicly released models include FlanT5 (Chung et al., 2022), OPT (Zhang et al., 2022), Llama 1 & 2 (Touvron et al., 2023a,b), Vicuna (Chiang et al., 2023), and more recently, Llama-3 (AI@Meta, 2024).

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 (Xu et al., 2015) and visual question answering (VQA) tasks (Antol et al., 2015; Das et al., 2017; Anderson et al., 2018). 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 pretrained LLMs, tapping into the knowledge accrued from extensive textual corpora. These work include Flamingo (Alayrac et al., 2022), OpenFlamingo (Awadalla et al., 2023), Palm-E (Driess et al., 2023), BLIP-2 (Li et al., 2023b), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023b), IDEFICS (Laurençon et al., 2023), MiniGPT-4 (Zhu et al., 2023) and many more (Li et al., 2023a; Ye et al., 2023; Gong et al., 2023; Gao et al., 2023; Zhang et al., 2023a; Su et al., 2023; Lyu et al., 2023), 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 (Liu et al., 2023b; Li et al., 2023c).

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 and human preference data, and (3) scaling the LLM parameters to $70B$ via an efficient pre-training approach.

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. During alignment training we freeze the parameters of the underlying LLM, allowing the projection layers to reach convergence faster than if trained end-to-end, and to inherit the reasoning capabilities of the LLM at inference time. To maximize feature compatibility between the modality encoders and the LLM, we use pre-trained encoders $g(\cdot)$ that have already been aligned to a text embeddings space, *e.g.* CLIP (Radford et al., 2021; Schuhmann et al., 2022) for images, CLAP (Wu* et al., 2023) for Audio signals, or IMU2CLIP (Moon et al., 2022) for IMU signals. For each text caption and modality pair $(\mathbf{X}_{\texttt{text}}, \mathbf{X}_{\texttt{MM}})$, we align them using the following objectives with a projection module (*i.e.* Perceiver Resampler (Alayrac et al., 2022) for vision encoder, and linear layers for other modalities).

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

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

To handle modalities larger than what can be accepted by the encoder $g(\cdot)$ (*e.g.* high-resolution images, long audio clips, *etc.*), we split the modality into pieces $\left(\mathbf{X}_{\texttt{MM}}^{[1]}, \mathbf{X}_{\texttt{MM}}^{[2]}, \ldots \mathbf{X}_{\texttt{MM}}^{[k]}\right)$ and project each piece independently, concatenating the result:

$$\mathbf{Z}_{\texttt{MM}}^{[i]} = \texttt{Projection}_\theta(h_{\texttt{latents}}, g(\mathbf{X}_{\texttt{MM}}^{[i]})) \quad (3)$$

$$\mathbf{Z}_{\texttt{MM}} = \mathbf{Z}_{\texttt{MM}}^{[1]} \parallel \mathbf{Z}_{\texttt{MM}}^{[2]} \parallel \ldots \parallel \mathbf{Z}_{\texttt{MM}}^{[k]} \quad (4)$$

Audio, IMU signals and videos are split into fixed-length pieces in the time dimension. For images, similar to Liu et al. (2024a), we split the image into an NxN grid after resizing to the next largest multiple of the encoder's input resolution. However, since we use a Perceiver Resampler to compress the image embeddings into a smaller number of tokens, we can use much larger grids for high-resolution images (up to 9.1 megapixels) without exceeding the LLM's maximum context length. The exact hyperparameters used at inference time are shown in Appendix E.3.

Training Optimization: Training a 70B model presents significant challenges due to memory usage limits during training. While quantization strategies (4 bits and 8 bits) (Dettmers et al., 2023) are popular choices, they often incur a trade-off between precision and accuracy at inference time.

To minimize GPU memory usage during training, we implement 3D parallelism using FSDP (Zhao et al., 2023) (for sharding model parameters, gradients, and optimizer states), interleaved tensor, and sequence parallelism (Korthikanti et al., 2022), and context parallelism (Liu et al., 2023a) for handling large sequences.

We provide more details on scaling the training pipeline in Appendix B.

3.2 Supervised 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. 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 (Hu et al., 2021) to further tune the LM behaviors.

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 beyond simple QA queries (*e.g.* "*Create a poem using this image*", "*Extract the phone number on this flyer*").

Therefore we collect 60K examples of high-quality multimodal instruction tuning data for multiple modalities using an iterative model in the loop process, as illustrated in Table 10 in Appendix C. Annotators are instructed to provide queries that are strictly multimodal, such that they cannot be

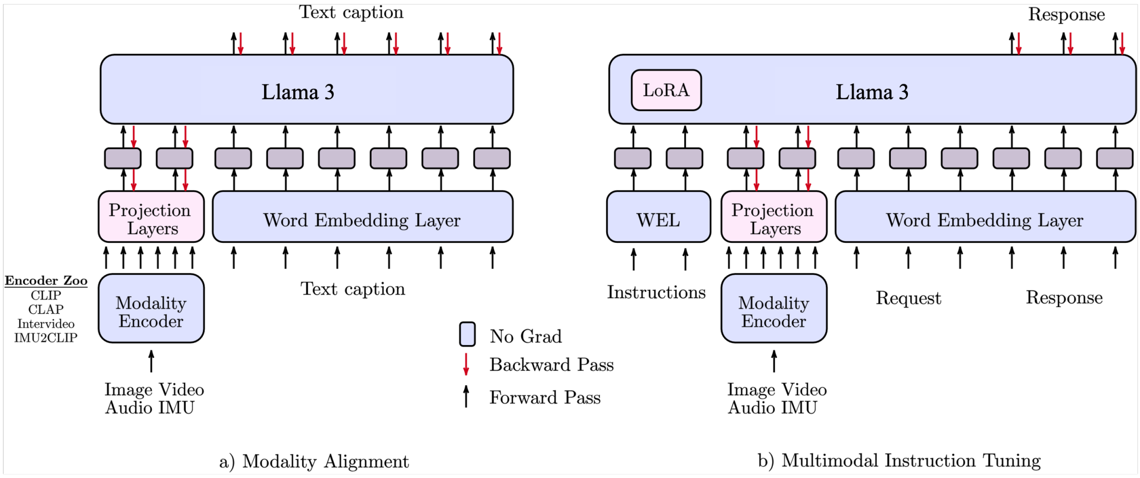


Figure 2: AnyMAL Training. (a) Modality alignment pre-training allows for mapping the output of each modality encoder into the joint LLM embeddings space through projection layers. (b) With multimodal instruction tuning, the model learns to associate system instructions and text queries with input multimodal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

answered without understanding the accompanying multimodal context. We then generate model responses using the queries and ask annotators to correct them as needed, which helps reduce annotation errors compared to having to construct responses from scratch.

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

Synthetic Augmentation: In addition to the high-quality ground-truth instruction tuning data above, we augment the dataset using the Llama-3 ($70B$) (AI@Meta, 2024) model, following similar approaches proposed by LLaVA (Liu et al., 2023b). 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 GPT-4.

3.3 Human Preference Alignment

Direct Preference Optimization: We further fine-tune the model on pairwise human preference data using Direct Preference Optimization (DPO) (Rafailov et al., 2023). Specifically, we initialize a policy π_θ and reference model $\pi_{\texttt{ref}}$ using the SFT'ed model. Given modality $\mathbf{X}_{\texttt{MM}}$, instruction $\mathbf{Z}_{\texttt{i}}$, preferred response $\mathbf{Z}_{\texttt{r}}^{+}$ and dispreferred response $\mathbf{Z}_{\texttt{r}}^{-}$, we optimize the loss:

$$\mathcal{L}_{\text{dpo}} = -\log \sigma \left(\beta \log \frac{r(\mathbf{X}_{\texttt{MM}}, \mathbf{X}_{\texttt{i}}, \mathbf{X}_{\texttt{r}}^{+})}{r(\mathbf{X}_{\texttt{MM}}, \mathbf{X}_{\texttt{i}}, \mathbf{X}_{\texttt{r}}^{-})} \right) \quad (5)$$

$$r(\mathbf{X}_{\texttt{MM}}, \mathbf{X}_{\texttt{i}}, \mathbf{X}_{\texttt{r}}) = \frac{p_\theta(\mathbf{X}_{\texttt{r}} | \mathbf{X}_{\texttt{MM}}, \mathbf{X}_{\texttt{i}})}{p_{\texttt{ref}}(\mathbf{X}_{\texttt{r}} | \mathbf{X}_{\texttt{MM}}, \mathbf{X}_{\texttt{i}})} \quad (6)$$

where β is the hyperparameter controlling the strength of the KL penalty. We train for 1 epoch on a dataset of $11K$ (image, query, preferred response, rejected response) tuples, where images and queries are sourced from the MM-IT dataset, and responses are generated using a number of models trained during the development process. Details of the DPO dataset collection are provided in Appendix C.

4 Experiments

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 and industry.

VQA Benchmarks: Table 1 shows the zero-shot performance on the MMMU dataset (Yue et al., 2024), VQAv2 (Antol et al., 2015), TextVQA (Singh et al., 2019), MMBench (Liu et al., 2024c), AI2D (Kembhavi et al., 2016), and MathVista (Lu

Models		MMMU	VQAv2	TextVQA	MMBench	AI2D	MathVista	ChartQA
OpenFlamingo	(Awadalla et al., 2023)	-	50.5	24.2	5.7	-	-	-
Flamingo-80B	(Alayrac et al., 2022)	-	56.3	35.0	-	-	-	-
InstructBLIP	(Dai et al., 2023)	-	-	50.7	33.9	-	-	-
IBELICS-80B	(Laurençon et al., 2023)	-	60.0	30.9	54.6	-	-	-
CogVLM	(Wang et al., 2023)	41.1	-	-	77.6	-	34.5	-
Llava-Next-34B	(Liu et al., 2024b)	51.1	-	69.5	79.3	-	46.5	-
InternVL 1.5-26B	(Chen et al., 2024)	45.2	-	-	-	80.7	53.5	**83.8**
Claude 3 Haiku	(Anthropic, 2024)	50.2	-	-	60.6	86.7	46.3	81.7
Gemini Pro	(Team et al., 2023)	47.9	71.2	73.5	75.2	73.9	52.1	74.1
Claude 3 Sonnet	(Anthropic, 2024)	53.1	-	-	67.8	88.7	47.9	81.1
Grok 1.5	(xAI, 2024)	53.6	-	78.1	-	88.3	52.8	76.1
Claude 3 Opus	(Anthropic, 2024)	59.4	-	-	63.9	88.1	50.5	80.8
GPT4V	(OpenAI, 2023)	56.8	77.2	78.0	81.4	78.2	49.9	78,5
Gemini Ultra	(Team et al., 2023)	59.4	77.8	**82.3**	-	79.5	53.0	80.8
AnyMAL 8B		44.2	71.0	62.9	66.2	47.8	26.7	-
AnyMAL 70B		**60.4**	**78.7**	77.0	**81.7**	**88.8**	**57.8**	81.7

Table 1: **Zero-shot Image-based QA** accuracy (%) results on 6 different VQA datasets (using pixels only, without external OCR model outputs). The top half of the baselines are the open-source models, whereas the bottom half are the proprietary models. **Bold** denote the top performance. AnyMAL demonstrates competitive zeroshot multimodal reasoning capabilities, compared to the baseline vision-language models.

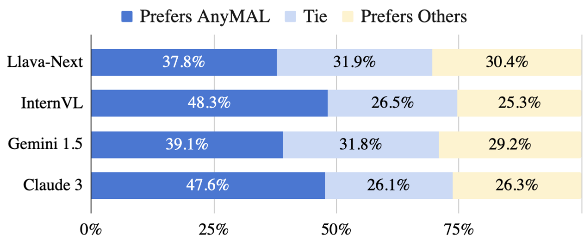


Figure 3: **Image-based reasoning pairwise human evaluation** results (% win, tie and lose) with baseline outputs *against* the AnyMAL responses on MM-IT (2K test set). AnyMAL responses are preferred by human judges more frequently than the baseline responses.

Models		MM-IT Accuracy
InternVL 1.5	(Chen et al., 2024)	66.5
Llava-Next	(Liu et al., 2024b)	67.8
Claude 3	(Anthropic, 2024)	62.2
Gemini 1.5	(Team et al., 2023)	65.0
AnyMAL 70B		**71.8**

Table 2: **Human evalaution of Image-based Reasoning**. We sample $2K$ multimodal queries each from MM-IT, and report the percentage of responses deemed by human annotators to be relevant to the query, factually correct and without any hallucinations.

et al., 2024) compared against the models in the literature that report zero-shot results on the respective benchmark. We focus on zero-shot evaluation to best estimate the model's performance on the open-ended queries at inference time.

Overall, our AnyMAL exhibits competitive performance compared to the industry-leading models (*e.g.* Gemini, GPT4) across multiple tasks, despite the relatively small number of parameters. Among the base LLM models for AnyMAL, 70B shows the most robust performance, underscoring the influence of substantial reasoning proficiency inherent in larger LLMs on tasks involving visual reasoning.

Human Evaluation on Image-based Reasoning Tasks: We evaluate the performance of our models against the most competitive vision-language models publicly available to run inference on (*i.e.* Gemini 1.5 Pro (Team et al., 2023), and Claude 3 Opus (Anthropic, 2024), LLaVA-NeXT (Liu et al., 2024b), InternVL (Chen et al., 2024)). Since the responses are subjective in nature (*e.g.* creative writing – "*Write a 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 $2K$ test set (Figure 3), as well as the pointwise evaluation (see the full rubrics in Appendix C.2). Specifically, we use the response accuracy which measures whether the response contains the relevant, factually correct and verifiable information (without any hallucinations) with regards to the image and the instruction.

Table 2 shows the pointwise evaluation on the

Models		EgoS	MVB
mPLUG-Owl	(Ye et al., 2023)	31.0	29.7
LongViViT	(Papalampidi et al., 2024)	33.0	-
VideoChatGPT	(Maaz et al., 2023)	-	32.7
VideoLlama	(Zhang et al., 2023b)	-	34.1
Gemini 1.5	(Team et al., 2023)	63.0	-
AnyMAL 70B		**66.8**	**46.4**

Table 3: **Zero-shot Video-based QA** accuracy (%) on **EgoS**chema, and **MVB**ench. AnyMAL demonstrates competitive zeroshot multimodal reasoning capabilities, compared to the baseline vision-language models.

Models		AudioCaps		
		CIDEr	SPICE	SPICEr
TD-Aligned	(Kim et al., 2019)	59.3	14.4	36.9
CNN10-VGG	(Xu et al., 2021)	66.0	16.8	41.4
ACT	(Mei et al., 2021)	67.9	16.0	42.0
PANNs + BERT	(Liu et al., 2022)	66.7	17.2	42.0
AnyMAL 7B (CLAP)		70.4	21.0	45.7
AnyMAL 13B (CLAP)		72.1	22.0	47.0
AnyMAL 70B (CLAP)		**77.8**	**23.0**	**50.4**

Table 4: **Zero-shot 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 understanding.

MM-IT test set. Specifically, it can be seen that AnyMAL attains the highest response accuracy and relevancy score (10.4% relative improvement compared to the strongest baseline: Gemini 1.5). This result highlights the enhanced capability of the model to comprehend and precisely answer questions in accordance with provided instructions. In Figure 3, we show that AnyMAL responses are preferred more frequently than the baseline model responses in the side-by-side pairwise evaluation, confirming the trend in the pointwise evaluation.

Video QA benchmarks: We evaluate our model on two challenging video question-answering benchmarks in Table 3: MVBench (Li et al., 2024), and EgoSchema (Mangalam et al., 2024). Our model demonstrates competitive results compared to the baselines, and achieves state-of-the-art performance. 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 (Yu et al., 2023). Our approach is compatible with such strategies, however that is beyond the scope of our experiments.

Audio Caption Generation: Table 4 shows the audio captioning results on the AudioCaps (Kim et al., 2019) 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. 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. Table 7 show example model outputs for audio reasoning tasks.

IMU Motion Description Generation: We use Ego4D (Grauman et al., 2022) to train an IMU-aligned AnyMAL, leveraging the synchronized IMU sensor data and textual narrations. 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, we achieve 52.5 CIDEr and 23.2 ROUGE-L against the ground-truth captions, showing the feasibility of the newly proposed task.

Qualitative Analysis: We provide example outputs from AnyMAL in Appendix A, and qualitative analysis against the baselines for each modality. Tables 5, 6 show outputs from various vision-language models on diverse example image and prompt pairs, compared with AnyMAL. Combining the audio and IMU captioning ability with the reasoning capbility of LLMs, in Tables 7, 9, and 8 we show examples of *novel* applications AnyMAL allows, *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), or interleaving multiple modalities (*i.e.* vision + IMU signals) for complex reasoning tasks.

5 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 LLMs within the multimodal settings.

References

AI@Meta. 2024. Llama 3 model card.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Anthropic. 2024. Introducing the next generation of claude.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. 2022. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Neftune: Noisy embeddings improve instruction finetuning.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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.

Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Reducing activation recomputation in large transformer models.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. 2023c. Mit: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023a. Ring attention with blockwise transformers for near-infinite context.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Xubo Liu, Xinhao Mei, Qiushi Huang, Jianyuan Sun, Jinzheng Zhao, Haohe Liu, Mark D Plumbley, Volkan Kilic, and Wenwu Wang. 2022. Leveraging pre-trained bert for audio captioning. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1145–1149. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player?

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2024. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36.

Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. 2021. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*.

Eric Mitchell. 2023. A note on dpo with noisy preferences & relationship to ipo. https://ericmitchell.ai/cdpo.pdf.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2022. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *arXiv preprint arXiv:2210.14395*.

NVIDIA. 2022. TransformerEngine. https://github.com/NVIDIA/TransformerEngine.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. 2024. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14386–14397.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6967–6977.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. 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*.

xAI. 2024. Grok-1.5 vision preview.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu. 2021. 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)*, pages 905–909. IEEE.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Qualitative Analysis

Image-based Reasoning: We provide qualitative examples in Tables 5, 6 comparing outputs from various baselines (Liu et al., 2024b; Chen et al., 2024; Team et al., 2023; Anthropic, 2024).

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

To keep the response concise, we add the following phrase to each query: "Keep your answers within 1-2 sentences unless necessary and do not exceed a maximum of 40 words."

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

Reasoning with IMU Motion Signals: Combining the IMU captioning ability with the reasoning capbility of LLMs, in Table 9 we show examples of *novel* applications AnyMAL allows, *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).

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 zero-shot capabilities of handling such interleaved modalities in Table 8 (*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.

B Training Optimization

To fully utilize the pre-trained language model's context length and limit the number of padding tokens used, we pack samples together in the sequence dimension and use a block diagonal attention bias to prevent the model from attending to hidden states across sample boundaries.

To limit the amount of GPU memory required to train the model, we implement 3D parallelism:

1. We use FSDP (Zhao et al., 2023) to shard model parameters, gradients and optimizer states.
2. We implement interleaved tensor and sequence parallelism as described by Korthikanti et al. (2022).
3. We shard the input sequence using context parallelism. (Liu et al., 2023a) to handle sequences that would not otherwise fit on a single tensor parallel group.

We store the sharded weights and gradients in fp32 and cast them to bf16 during the forward and backward pass using FSDP's mixed precision API. We use Transformer Engine (NVIDIA, 2022) fused modules, which increases bf16 training throughput by ~40%.

We reduce the number of activations recomputed during the backward pass by utilizing a custom backward method to implement activation checkpointing, following Touvron et al. (2023a).

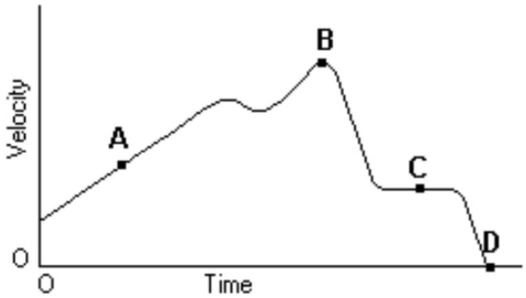
C Dataset Collection and Human Evaluation

C.1 Fine-tuning Dataset Collection

SFT Dataset Collection: We collect 60K examples of high-quality multimodal instruction tuning data for multiple modalities, as illustrated in Table 10. Specifically, we use various Creative Commons licensed, publicly available images, augmenting these images with manually created instructions and responses.

DPO Data Collection: We train on a dataset of 11k (image, query, preferred response, rejected response) tuples. Response pairs were initially generated by using two variants with and without SFT tuning. We update our response pairs further through iterative training until evaluation metrics stabilize.

Example 1. Science Understanding

Input Image	
Prompt	***Given the graph of the velocity vs. time of a duck flying due south for the winter. at what point did the duck stop its forward motion?***
InternVL	The duck stopped its forward motion at point C on the graph, where the velocity is zero.
LLaVA-NeXT	The duck stopped its forward motion at point C on the graph.
Gemini 1.5	The duck stopped its forward motion at point D where the velocity reaches zero.
Claude	The duck stopped its forward motion at point C. This is where the velocity reaches zero on the graph, indicating the duck came to a stop before starting to move backwards.
AnyMAL 70B	The duck stopped its forward motion at point D.

Example 2. Landmarks

Input Image	
Prompt	***Where is this place located? also, what is the name of this place?***
InternVL	The image shows a large stadium with a dome-shaped roof, and it is located in the United States. The name of the place is not visible in the image.
LLaVA-NeXT	This place is located in the United States, and it is the Metrodome, a stadium in Minneapolis, Minnesota
Gemini 1.5	This is the Kingdome, a multi-purpose stadium located in Seattle, Washington.
Claude	This image shows the Louisiana Superdome, a large multi-purpose stadium located in New Orleans, Louisiana. It is home to the New Orleans Saints NFL team.
AnyMAL 70B	This is the Kingdome, a former multi-purpose stadium located in Seattle, Washington.

Table 5: Examples comparing AnyMAL with various baselines. AnyMAL demonstrates competitive performance in science and less popular landmarks

Example 3. Historical Facts

Input Image	
Prompt	***When was the first time one of these desserts were made?***
InternVL	The first time a chocolate cake with cherries was made is not discernible from the image.
LLaVA-NeXT	The first chocolate cake with cherries on top was made in the 19th century, but the exact date is not known.
Gemini 1.5	There is no information about when these desserts were first made.
Claude	The first known chocolate cake recipe dates back to 1764, when it appeared in Dr. James Baker's book "The Experienced English Housekeeper". However, cakes in general have a much longer history, with ancient Egyptian bread sweetened with honey being an early precursor.
AnyMAL 70B	The first chocolate cake was made in 1764 by Dr. James Baker.

Example 4. Recipe

Input Image	
Prompt	***what is a recipe for a cocktail that would resemble the one shown here.***
InternVL	To make a cocktail similar to the one in the image, you would need a champagne flute, a bottle of champagne, and a slice of lemon. Fill the champagne flute with champagne, add a slice of lemon on the rim, and enjoy.
LLaVA-NeXT	A cocktail that resembles the one shown in the image could be a simple glass of sparkling wine with a twist of lemon. To make it, pour a glass of sparkling wine, such as champagne or prosecco, and add a twist of lemon for a refreshing touch.
Gemini 1.5	The cocktail in the image resembles a **French 75**. It is made with gin, lemon juice, simple syrup, and topped with Champagne.
Claude	The cocktail resembles a **French 75**, which is made by combining gin, lemon juice, simple syrup, and champagne or sparkling wine, then garnishing with a lemon twist
AnyMAL 70B	A recipe for a cocktail that resembles the one shown here is a **French 75**, made with gin, lemon juice, simple syrup, and champagne.

Table 6: Examples comparing AnyMAL with various baselines. AnyMAL demonstrates competitive performance in both historical facts (*e.g.* Chocolate cake creator) and identifying objects to share recipes(only the bigger models got the drink right).

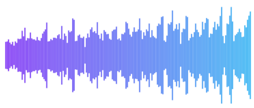
Audio Signals Input	AnyMAL Output
	**Gold Caption**: Typing on a keyboard is occurring in a quiet environment **Prompt**: Can you guess my location? Reply with one or two sentences. **Response**: Based on the sound of the typing, I would guess that you are located in a quiet, indoor setting, possibly 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.
	**Gold Caption**: Very strong wind is blowing, and waves are rustling **Prompt**: Based on the weather, what should I wear? Reply with one or two sentences. **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.
	**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. **Prompt**: Describe the sound in one or two sentences. **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.

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.

C.2 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).

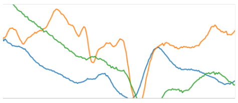
Input Modality 1	Input Modality 2	AnyMAL Output
 (biking.npy)		**Prompt**: (Given the motion signals) Write a social media caption for this view. **Response**: Pedaling along the San Francisco Bay, taking in breathtaking views of the Golden Gate Bridge! The salty sea breeze invigorates me as I make my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

Table 8: 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.

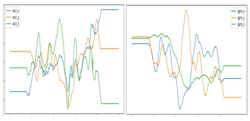
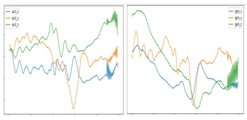
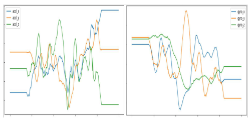
IMU Signals Input	Reference Video	AnyMAL Output
		**Prompt**: Describe the motion. **Response**: [Camera wearer] looks around.
		**Prompt**: Describe the motion. **Response**: [Camera wearer] walks in the house.
		**Prompt**: What's the safest way to stop? **Response**: To stop safely on a bike, use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

Table 9: 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.

D 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 & output texts, and (3) multimodal combination of input images and text outputs. If any of these detect an integrity violation, we proceed to reject the entire user query. For streaming use-cases, we run output classifiers for each generated sentence and halt generation if any detect an integrity violation.

(1) Input image: we use a pre-trained image classifier based on RegNetY (Radosavovic et al., 2020) to detect any content that violates integrity standards. This detection encompasses graphic material, violent imagery, hate symbols, instances of bullying, harassment, *etc*.

(2) Input & output texts: we use a RoBERTa-based text classifier (Liu et al., 2019) trained to detect integrity-violating utterances such as violence, harassment, hate speech, *etc*. within the input text or generated output text.

(3) Multimodal association in input image & output text: we use a multimodal AnyMAL-based classifier conditioned on the input image and generated response to account for cases where individually-benign images and texts might be integrity-violating when associated.

Training Time Safety: The datasets used for pre-

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 10: 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.

training (e.g. (Radenovic et al., 2023; Singer et al., 2022)) 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-3 (the version we report most of our results on) places safeguards such as negative example fine-tuning, reinforcement learning with human feedback (RLHF) (Christiano et al., 2017; Bai et al., 2022; Rafailov et al., 2023).

E Additional Notes on Experiments

E.1 Multimodal Prompts

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.

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: <img>*". 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: <img>*". 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: <img>*". User message: "*In the image, {question} Reply in one prahse/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: <img>\n This meme contains text: '{ocr}'*". 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: <img>*". 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."}.*"

MMMU, ChartQA, AI2D 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: <img>.*" *User message: "{context} Question: {question} \n\n Options: {choices} \n\n Reply in a single letter.*"

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.*

EgoSchema, MVBench 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.*"

E.2 Multimodal Inputs

Figure 4 shows the diagram for performing modality-interleaved inference (for examples shown in Table 8).

E.3 Hyperparameters

Pre-training: Table 11 report the hyperparameters used in this work for model pre-training.

Supervised Fine-tuning: We use LoRA adapters to fine-tune the projection layers and language model on the MM-IT training set with the prompt described in E.1. We initialize the projection layer

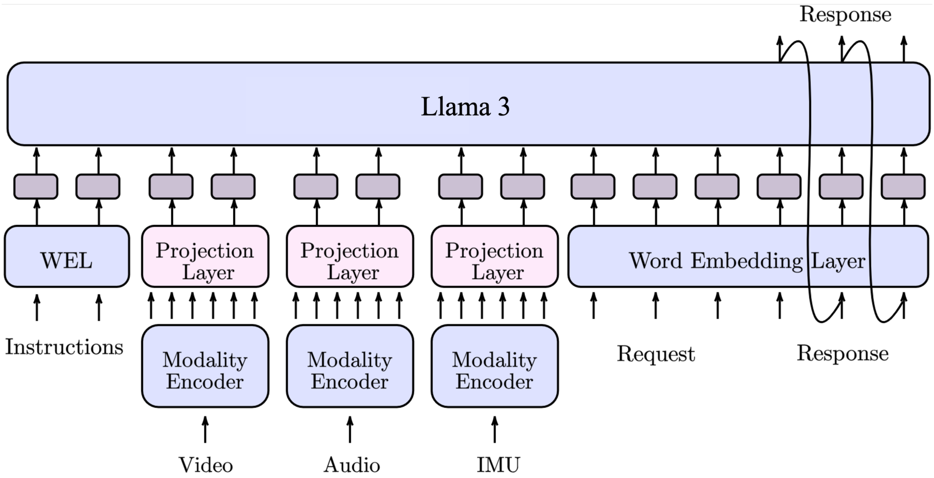


Figure 4: AnyMAL Inference example with multiple modality as input.

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)	26880	2×10^{-4}	25k	64	Resampler (12)
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 (8B, IMU)	256	1×10^{-4}	2k	32	Linear (1)

Table 11: Hyperparameters for AnyMAL Pre-training

using the weights produced by the pre-training process. We set LoRA $r = 8$, $\alpha = 32$, and add LoRA modules on all linear layers. We finetune the model for 3k steps with a global batch size 128. Training warms up to an initial learning rate of 5×10^{-5} linearly over 40 steps, decaying to 10% of the initial learning rate via a cosine schedule by the end of training. We apply NEFTune (Jain et al., 2023) to the language model with $\alpha = 10$.

Human Preference Alignment: We initialize the reference model and policy using the Supervised Fine-tuned model. For the policy, we continue fine-tuning the LoRA adapters that we trained during Supervised Fine-tuning, and keep all other parameters frozen. We use a global batch size of 32. Training warms up to an initial learning rate of 1×10^{-5} linearly over 20 steps, decaying linearly to 2/3 of the initial learning rate by the end of training. We use Conservative DPO (Mitchell, 2023) with the label smoothing parameter $\epsilon = 0.05$. We apply NEFTune (Jain et al., 2023) to the language model with $\alpha = 10$.

E.4 Code Base & Hardware

The implementations of the transformer-based models are extended from the HuggingFace[1] code base (Wolf et al., 2020) and other cited authors' released code-bases. Our entire code-base is implemented in PyTorch (Paszke et al., 2019). All models in this work are trained on a varying number of Nvidia A100 [2] and H100 [3] GPUs.

F 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.

[1]https://github.com/huggingface/transformers
[2]https://www.nvidia.com/en-us/data-center/a100/
[3]https://www.nvidia.com/en-us/data-center/h100/

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.