MLAN: Language-Based Instruction Tuning Preserves and Transfers Knowledge in Multimodal Language Models

Jianhong Tu[1]*, **Zhuohao Ni**[2]*, **Nicholas Crispino**[1], **Zihao Yu**[1], **Michael Bendersky**[3], **Beliz Gunel**[3], **Ruoxi Jia**[4], **Xin Liu**[5], **Lingjuan Lyu**[6], **Dawn Song**[7], **Chenguang Wang**[1]†

[1]Washington University in St. Louis [2]The University of British Columbia
[3]Google Research [4]Virginia Tech [5]University of California, Davis
[6]Sony AI [7]University of California, Berkeley

{jianhong.t, ncrispino, yu.zihao, chenguangwang}@wustl.edu
peterni@student.ubc.ca {bemike, bgunel}@google.com ruoxijia@vt.edu
xinliu@ucdavis.edu Lingjuan.Lv@sony.com dawnsong@berkeley.edu

Abstract

We present a novel visual instruction tuning strategy to improve the zero-shot task generalization of multimodal large language models by building a firm text-only knowledge base. Existing work lacks sufficient experimentation on the importance of each modality in the instruction tuning stage, often using a majority of vision-language data while keeping text-only data limited and fixing mixtures of modalities. By incorporating diverse text-only data in the visual instruction tuning stage, we vary vision-language data in various controlled experiments to investigate the importance of modality in visual instruction tuning. Our comprehensive evaluation shows that the text-heavy instruction tuning approach is able to perform on-par with traditional vision-heavy mixtures on both modalities across 12 general datasets while using as low as half the total training tokens. We find that simply increasing sufficiently diverse text-only data enables transfer of instruction following ability and domain knowledge across modalities while being more efficient than the vision-language approach.

1 Introduction

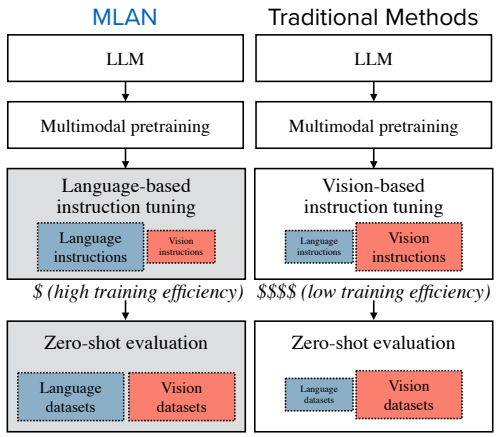
Multimodal large language models (MLLMs) have advanced and enabled a wide range of vision-language tasks such as visual question answering and image captioning (Liu et al., 2023b; Alayrac et al., 2022; Li et al., 2023b; Lin et al., 2023; Bai et al., 2025). Their zero-shot generalization ability to unseen tasks has the potential to further revolutionize broader real-world applications (Driess et al., 2023; Zhu et al., 2023; Li et al., 2023a). To construct MLLMs, vision-language pretraining is performed on a large scale with image-text data, aligning the modalities before visual instruction tuning aligns the model with human preferences (Liu et al., 2023a; Dai et al., 2024; Lin et al., 2023). The importance of strong vision-language pretraining is established, with more data resulting in greater improvements in instruction-following abilities and downstream performance (McKinzie et al., 2024; Zhang et al., 2024a). However, current visual instruction tuning practices overwhelmingly rely on image-text pairs and large-scale vision-language datasets. This emphasis introduces a significant distributional shift from the language-rich corpora used during pretraining, often degrading the model's general language understanding and leading to catastrophic forgetting of core knowledge (Zhang et al., 2024b). Given the similarity in instruction tuning data across modalities and the strong modality alignment achieved with vision-language pretraining, we believe text-only data is underutilized in existing training mixtures. Additionally, various design choices regarding the instruction tuning dataset composition with respect to modalities are underexplored.

In this work, we introduce **MLAN** (**M**ultimodal **LAN**guage-based instruction tuning), a new perspective in vision instruction tuning that treats language as the primary way to unlock knowledge during instruction tuning (Figure 1). Our key insight is that instruction-following abilities and domain knowledge, once acquired through diverse language-only tasks, can generalize across modalities with minimal vision-language supervision. By grounding vision capabilities in a small number of targeted image-text examples, we maintain high performance across both vision and text tasks while significantly reducing training costs. Specifically, with MLAN we unlock vision instruction following abilities by teaching a pre-trained model to execute text-only instructions and then complementing the dataset with a relatively small portion of vision-language examples in a domain adaptation fashion.

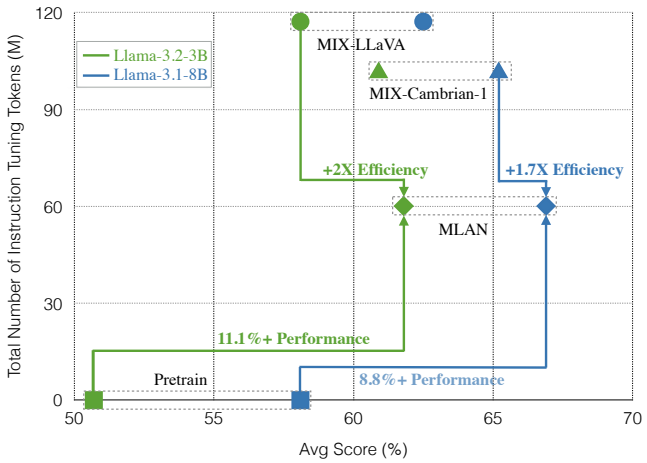
To demonstrate MLAN's effectiveness, we pre-

* Equal contribution
† Corresponding author

Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM), pages 59–74



(a) Comparison of MLAN with standard visual instruction tuning.



(b) Main results on evaluation tasks, averaged over text-only and vision-language performance.

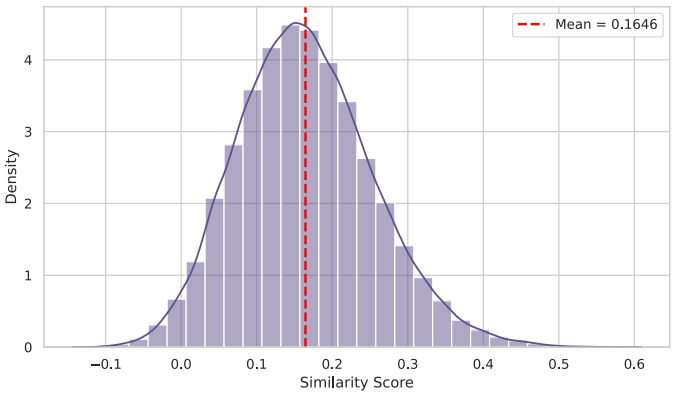
Figure 1: Overview of MLAN. (a) MLAN represents a shift in perspective towards text during instruction tuning. After vision-language pretraining, we include diverse text-only data in our instruction tuning mixture spanning many tasks. We emphasize including text-only data to show the transferability of instruction tuning across modalities. For evaluation, we select ample text-only and vision-language datasets, allowing us to compare performance changes across modalities. (b) We evaluate MLAN on two pretrained multimodal models based on Llama-3.2-3B and Llama-3.1-8B across unseen language and vision benchmarks, achieving comparable performance at higher training efficiency (up to almost 2x as efficient compared to standard vision-heavy instruction tuning) with our language-based approach.

train MLLMs over a variety of settings based on Llama-3.2-3B (Meta AI, 2024) and Llama-3.1-8B (Dubey et al., 2024), following the state of the art multimodal training mechanism (Liu et al., 2023b,a), varying only the dataset. We then apply MLAN to the MLLMs and observe the following key insights over both models on average during evaluation on 12 comprehensive benchmarks across language and vision modalities. (1) Compared with the traditional vision-heavy finetuning approaches of LLaVA (Liu et al., 2023a) and Cambrian-1 (Tong et al., 2024), our models finetuned with MLAN demonstrate a matching or better performance on downstream vision-language tasks while seeing less than half of the images and consistently showing better text-only performance. We show that text-only data is imperative to obtain world knowledge and understanding of complex instructions, even in the vision domain. (2) Text-only instruction tuning is more cost-effective. The rich and dense information compensates for the limited diversity in public vision datasets, allowing for superior performance while reducing the total number of processed training tokens by half. (3) Neither language nor vision alone is enough for a generalist MLLM. Our experiments show that while instruction following abilities may transfer across modalities, their impact on the other modality is limited: certain vision-language tasks do not benefit from text-only tuning and vision-language tuning can result in severe degradation of language abilities. However, mixing bi-modal data, even at a small percentage, leads to surprising performance boosts and achieves the best results in both modalities. We hope our findings will foster future research on language-centered training and instruction tuning, paving the way for fundamental advancements in large MLLMs.

2 Approach

MLAN views vision instruction following abilities as a natural extension of text-only abilities, a transfer that can occur due to the extensive multimodal pretraining used in MLLMs.

We begin by motivating our method through an empirical analysis of similarities in text-only and vision-language instruction tuning data, which leads to our hypothesis that text-only data can largely replace vision-language data to improve performance on general tasks. Then, we detail our training, following the standard design of existing instruction tuning methods (Wei et al., 2022; Xu et al., 2023; Dai et al., 2023; Liu et al., 2023b) in four stages: selecting training data, format-



(a) Distribution of cross-modal similarity scores between modalities with a non-negative mean by z-test ($p < 0.001$).

Vision-Language Instructions

INSTR 1: Your task involves classifying object images into their respective categories like Bed, Sink, Sneakers, Table, TV and so on...

INSTR N: Each image has something going on. Carefully analyze the image and generate 5 captions for each image.

CONTEXT: <image>

OUTPUT: <text>

Text-Only Instructions

INSTR 1: Given a text passage... your task is to classify the item being sold into exactly one of these categories: 'housing', 'furniture', 'bike', 'phone', 'car', 'electronics'...

INSTR N: In this task, you are given a conversation, and your task is to generate a summary from the information present in the given conversation...

CONTEXT: <text>

OUTPUT: <text>

(b) Examples of instructions across modalities that share similar goals.

Figure 2: Similarity between text-only and vision-language instruction tuning data shown both (a) quantitatively with similarity scores and (b) qualitatively with examples. 100k instructions are sampled from the Super-NaturalInstructions (Wang et al., 2022b) and Vision-Flan (Xu et al., 2024) datasets and embedded by a pretrained sentenceTransformer, all-mpnet-base-v2 (Song et al., 2020). The red vertical line denotes the mean score. We then randomly sample and display two instructions with high cosine similarities (0.53 & 0.38).

ting the data with instructions, fine-tuning a pretrained MLLM on the training set (Sec. 2.2), and evaluating the instruction tuned model on standard academic benchmarks in the zero-shot setting (Sec. 2.3).

2.1 Natural Correspondence between Text-Only and Vision-Language Instructions

While the image-text and the text-only distribution of instructions significantly differ from each other, we observe shared semantics and structure on the task level when comparing wild instruction-response pairs in both modalities.

Semantic Similarity We study two comprehensive large-scale instruction tuning datasets with one from each modality, namely Super-NaturalInstructions (Wang et al., 2022b) and Vision-Flan (Xu et al., 2024), which are representative of common structures and tasks. We show vision-language and text-only tasks are similar by randomly sampling 100k instances from each dataset and examining the distribution of the cosine similarities between embedded instructions as shown in Figure 2(a). A significantly non-negative mean cosine distance provides evidence that the tasks performed in either domain are somewhat similar, based on the belief that tasks are defined by the instructions. Additionally, there is a small yet nonzero chance to even see a pair of tasks that are comparable with high similarities (>0.3) in the language and vision domain. To qualitatively demonstrate this, in Figure 2(b) we show two pairs of semantically similar instructions from each datasets with a similarity score of 0.53 and 0.38, respectively. While the first example is a classical classification task, the second requests a concise representation of the context, where the context may be a text paragraph or an image. We reason that if the ability to describe a casual conversation is acquired, the ability to caption an image can be readily obtained.

Structural Similarity The well-established problem of solving zero-shot tasks can be split into a user prompt followed by a model's response for both modalities. While some text-only tasks appeal to a model's internal knowledge, such as ARC (Bhakthavatsalam et al., 2021), the task of open-book question answering is analogous to vision question answering in the sense that additional inputs are provided to serve as the reference where the final answer is derived. If the vision and the text modalities are well aligned, it makes sense for a model to easily refer to the details in an image as the image tokens are no different than the native word tokens in its embedding space.

2.2 Training Details

Our approach, MLAN, is simple, changing the dataset composition across modalities compared to traditional MLLM instruction tuning. We fine-tune a multimodal pretrained LLM in the FLAN-style (Wei et al., 2022) and further train on a small portion of vision instruction data (compared to the

number of text-only instances) to adapt the model to vision-language queries. While mainstream methods, including LLaVA (Liu et al., 2023b) and Cambrian-1 (Tong et al., 2024), also include some text-only examples in their vision instruction tuning dataset, their primary goal has been providing language as a form of regularization to prevent catastrophic forgetting. Our method differs by approaching vision instruction tuning from the other way around: we build strong language-only instruction-following abilities to build a robust knowledge base, and then introduce a small number of vision instances solely for grounding and domain adaptation. To demonstrate that adjusting the data composition alone is a viable substitute for vision-heavy instruction tuning, we use a fixed size budget and shared data sources for all our experiments, thus controlling the effect of longer training sessions and variable data quality.

Dataset Selection Inspired by the similarity in instruction tuning across modalities, we use the same two diverse datasets to train with, encompassing a multitude of tasks in each modality. For text-only data we sample from the over 1600 tasks in Super-NaturalInstructions (Wang et al., 2022b), while for vision-language data we sample from the 187 tasks in Vision-Flan (Xu et al., 2024). This gives us ample coverage across many text-only and vision-language tasks. For all of our experiments, we use a fixed data budget of 186,000 instances, which can come from either Super-NaturalInstructions or Vision-Flan depending on the setting.

Models and Multimodal Pretraining We follow the architecture design of LLaVA (Liu et al., 2023a) that connects a visual encoder with a projector that enables the LLM to use the outputs of the visual encoder to process image inputs in addition to texts. We choose CLIP-ViT-L/14@336 (Radford et al., 2021) and a two-layer MLP with GELU activation as the visual encoder and the projector, respectively. We select the base LLMs as Llama-3.2-3B (Meta AI, 2024) and Llama-3.1-8B (Dubey et al., 2024), both the non instruct versions. We conduct multimodal pretraining for both models on LLaVA-Pretrain-558K using the same hyperparameters as in Liu et al. (2023a). These models are then finetuned on our language-heavy training dataset for one epoch using a global batch size of 128, a cosine learning schedule, a learning rate of 2e-5, a warm-up ratio of 0.03, and no weight decay. Both the visual encoder and LLM are frozen throughout the pretraining session while the parameters in the MLP projector are updated. After pretraining, the visual encoder and the projector function as a visual tokenizer that turns an image into tokens compatible with the LLM.

Instruction Tuning To test our instruction tuning methodology, we finetune MLLM checkpoints using a controlled mixture of text-only and vision-language data, focusing on the former. This is because language, rather than vision, remains the primary medium for users to interact with models when they specify their needs. In contrast, most existing multimodal instruction tuning approaches prioritize vision-language data and include language-only tasks merely to mitigate forgetting. (Liu et al., 2023a; Bai et al., 2023; Ye et al., 2023; Luo et al., 2024; Tong et al., 2024). These approaches require many more training tokens and rely on a greater number of vision-language datasets. See Table 8 in Appendix D.1 for the percentage of text-only data included during instruction tuning for various state of the art MLLMs. Current instruction tuning mixtures across models vary substantially in language content, yet few of these design choices are grounded in systematic empirical comparison. Our method systematically tests the effectiveness of the composition of instruction tuning data by modality, then anchors in a shift in perspective, treating **language as the foundation** in instruction tuning.

2.3 Evaluation Tasks

Our evaluation suite covers diverse text-only and vision-language tasks for zero-shot evaluation that are not seen during training. The text-only benchmarks include **Commonsense understanding**, **Reasoning**, **Reading comprehension** and **Scientific knowledge testing**. Similarly, the selected vision-language benchmarks primarily test **Scene Understanding** and **Image Reasoning**. Notably, MMLU (Hendrycks et al., 2020), MMMU (Yue et al., 2024), and MME (Fu et al., 2023) are large multidisciplinary benchmarks covering wide domains. We craft suitable instruction templates for each dataset in the same way as for the training datasets, using the same collection of instruction prompts. The final evaluation collection includes 7 text-only datasets and 5 vision-language datasets. The answer types cover short-response, multiple-choice, and true/false questions. Appendix C provides a brief description of each dataset.

Models	Method	Vision Benchmarks					
		POPE	ScienceQA-IMG	MMMU	MME	MMBench	Avg.
Llama-3.2-3B	Pretrain	66.67*	43.73	26.44	700*	51.10	42.59
	MIX-LLaVA-1.5	80.10	64.65	29.00	1293.56	**67.71**	57.53
	MIX-Cambrian-1	81.90	**65.94**	28.67	1367.38	67.48	58.57
	MLAN	**83.17**	**65.94**	**29.33**	**1405.53**	67.01	**59.13**
Llama-3.1-8B	Pretrain	66.67*	63.81	27.67	700*	62.81	49.61
	MIX-LLaVA-1.5	79.90	67.97	30.89	1354.52	70.29	59.49
	MIX-Cambrian-1	**82.57**	70.55	**36.00**	1408.02	**73.50**	**62.58**
	MLAN	81.84	**71.15**	34.44	**1436.83**	72.51	62.25

Table 1: Zero-shot results on the held-out vision-language datasets for Llama-3.2-3B and Llama-3.1-8B. We compare Pretrain, MIX-LLaVA-1.5, MIX-Cambrian-1, and MLAN (ours). * denotes that the pre-trained models fail to generate meaningful responses other than all "yes" or "no". ScienceQA (Lu et al., 2022) is included in Vision-Flan but excluded in experiments. The MME scores are normalized by dividing by the maximum value (2800) when computing the average.

Models	Method	Language Benchmarks								
		ARC-E	ARC-C	CommensenseQA	PIQA	RACE	BoolQ	CosmosQA	MMLU	Avg.
Llama-3.2-3B	Pretrain	62.42	42.41	63.72	76.77	**70.37**	62.91	**67.77**	24.09	58.81
	MIX-LLaVA-1.5	69.40	43.34	58.39	78.40	58.57	68.93	47.57	44.65	58.66
	MIX-Cambrian-1	**71.68**	46.25	60.85	**79.27**	67.98	**71.59**	59.40	48.39	63.18
	MLAN	71.30	**46.93**	**66.18**	79.11	70.27	68.44	64.76	**49.03**	**64.50**
Llama-3.1-8B	Pretrain	71.09	50.00	70.19	80.14	79.41	64.89	76.65	39.79	66.52
	MIX-LLaVA-1.5	72.60	48.81	66.20	79.43	71.44	75.38	59.53	50.51	65.49
	MIX-Cambrian-1	72.80	48.81	68.88	80.03	74.22	77.22	64.42	55.69	67.76
	MLAN	**74.79**	**50.17**	**73.05**	**81.23**	**79.91**	**78.53**	**76.68**	**58.18**	**71.57**

Table 2: Zero-shot results on the held-out text-only datasets for Llama-3.2-3B and Llama-3.1-8B. We compare Pretrain, MIX-LLaVA-1.5, MIX-Cambrian-1, and MLAN (ours).

3 Experiments

In this section, we show that MLAN is both more effective and training efficient compared to the pre-trained MLLMs as well as state of the art multi-modal instruction tuning mixtures across all the tasks we evaluate. Additional details of the training and experimental setup are described in Appendix B.

3.1 Main Results

We compare various instruction tuning methods built upon our multimodal pretrained Llama-3.2-3B and Llama-3.1-8B. We include the following settings, all using our specified training methodology, only varying composition: (1) Pretrain: The MLLM after multimodal pretraining with no instruction tuning. (2) MIX-LLaVA-1.5 and MIX-Cambrian-1: We use our training dataset along with the multimodal instruction tuning mixture recipes of LLaVA (Liu et al., 2023a) and Cambrian-1 (Tong et al., 2024), i.e., with 6% and 25% text-only instruction data, respectively. (3) MLAN: Our text-first instruction tuning method with a composition heavily favoring (75%) text-only data.

Cross-Task Generalization We report the scores of pretrained MLLM and instruction tuned models on 12 benchmarks in Tables 1 and 2, respectively. Compared with MIX-Cambrian-1, MLAN yields the best performance in the 3B setting and matches the best score in the 8B setting, only falling behind by 0.33%, despite being trained on less than half of the images. The competitive vision performance shows effective cross-modal transfer. MLAN consistently improves performance on knowledge-intensive tasks such as MMLU, CosmosQA, and ARC-C, demonstrating stronger internal knowledge retention compared to vision-heavy baselines.

Knowledge Erosion We note that both MIX-LLaVA-1.5 and MIX-Cambrian-1 suffer from catastrophic forgetting, especially on CommonsenseQA (Talmor et al., 2019) and CosmosQA (Huang et al., 2019), showing performance degradations up to 5.3% and 20.2%. However, MLAN is more resilient against forgetting. In the only case where its performance decreases in CosmosQA, the decline is significantly smaller

than other models (3.1% vs. 20.2% & 8.37%). On all other benchmarks, including vision, our method shows a solid positive gain. Such an observation unveils an asymmetrical interaction between vision and text modalities, where the text ability is more susceptible to forgetting, but the vision ability generally benefits from language-based tuning. This trend is explored again in Section 3.2.

Method	Number of Tokens
Text-Only IT	37,906,142
MLAN	60,112,680
MIX-Cambrian-1	101,480,339↑68.8%
MIX-LLaVA-1.5	117,220,955↑95.0%
Full Vision-Language IT	122,054,758↑103.0%

Table 3: All token counts for various training settings with 186,000 total instances. The percentage score indicates the size increase relative to the MLAN setting.

3.2 Training Efficiency

A major advantage of our method is that it significantly reduces the computational cost measured by the number of training tokens processed by the base LLM compared to vision-based instruction tuning. Table 3 details the number of training tokens, including those in the visual prefix. Visual inputs drastically increase the training burden as an image is converted to hundreds of visual tokens (576 tokens with CLIP-ViT-Large-patch14@336 (Radford et al., 2021)) before being processed along with regular text tokens. Therefore, MLAN stands out as a more efficient vision instruction-tuning approach that avoids excessive instruction tuning on images.

3.3 Knowledge Transfer Curve

To better understand the role of language, we perform a controlled study by varying the proportion of language-only data in the instruction tuning mixture, increasing it in 12.5% increments. We show the performance of Llama-3.2-3B-based MLLMs with different amounts of language instruction data in Figure 3. Notably, we observe that even a small amount of language data (12.5%) leads to a sharp increase in both text and vision performance, suggesting that foundational knowledge acquired through language tuning quickly transfers across modalities. As the proportion of language data increases further, text performance continues to improve, whereas vision performance peaks and then slightly declines. Full vision-language tuning fails to match the peak vision performance achieved with a balanced mix, indicating that language-based knowledge is not only transferable but also essential for efficient vision instruction tuning. This analysis reinforces our central claim: language acts as a scaffold for multimodal reasoning, and a moderate inclusion of vision data is sufficient for grounding.

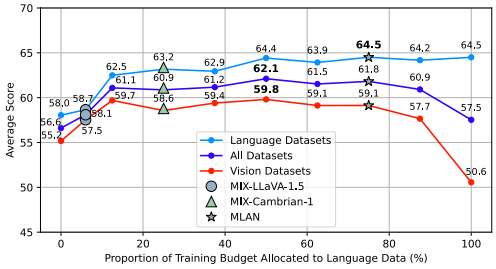


Figure 3: Average scores on Llama-3.2-3B based MLLMs with respect to the percentage of language data mixed in. The percentage denotes the amount of language data.

Base LLM	Variant	Text Avg.	Vision Avg.
Llama-3.2-3B	+MLAN	64.50	59.13
	+Instruct LLM	**67.74**	**60.98**
	-25% tasks	65.68	55.71
	-50% tasks	65.98	55.73
	-75% tasks	66.35	56.79

Table 4: Ablation study on Llama-3.2-3B with different instruction tuning variants and fewer tasks.

3.4 Additional Instruction Tuning Factors

Instruction Tuned Base Models We use base (non instruction tuned) models in our experiments to show the impact of text-only data while controlling the amount of text instruction tuning. However, mainstream vision instruction tuning methods mostly choose instruction-tuned (chat) models as the default LLM backbone (Liu et al., 2023b; Dai et al., 2024; Lin et al., 2023). Table 4 shows that finetuning the instruction tuned variant instead of the pretrained model readily boosts both text and vision performance by 2-3%, even when we continue to emphasize text-only data in the visual instruction tuning phase. This provides more evidence that the text-first approach throughout training is beneficial. A possible explanation for this is that the model adapts to the instruction following format and eliminates the distributional shift from the pretraining to the instruction tuning corpus.

Task Diversity within Datasets Prior work has emphasized the importance of diversity within in-

struction tuning datasets (Li et al.; Xu et al., 2024; Wei et al., 2022). We conduct a controlled finetuning experiment by reducing the proportion of included tasks (25%, 50%, 75%, 100%) while keeping the total number of training instances fixed. Surprisingly, text performance slightly improves with fewer tasks, peaking at 25%, while vision performance only improves with full task coverage. This suggests that task diversity does not uniformly benefit all modalities: some tasks may be less helpful, and that over-diversification may dilute useful supervision, especially for language.

Base LLM	PT	IT	Text Avg.	Vision Avg.
Llama-3.2-3B	LLaVA	Vision-Flan	55.14	57.61
		Super-Natural	**64.89**	46.95
	ShareGPT4V	Vision-Flan	58.05	**58.26**
		Super-Natural	64.48	50.20
Llama-3.1-8B	LLaVA	Vision Flan	63.08	54.77
		Super-Natural	**72.05**	52.55
	ShareGPT4V	Vision Flan	58.27	**56.84**
		Super-Natural	71.76	50.76

Table 5: Average performance across different vision pretraining (PT) and instruction tuning (IT) strategies.

3.5 Interaction between Pretraining and Single-Modal Instruction Tuning

Before visual instruction tuning, the vision pretraining step aims to align the text and vision modalities. Increasing pretraining data has been shown to increase post instruction tuning performance given the same corpus (McKinzie et al., 2024), but changes in pretraining data have been shown to have minimal effects (Cocchi et al., 2025). To investigate how the pretraining dataset affects instruction tuning on various modalities, we conduct experiments using single-modality instruction tuning datasets on another pretraining dataset (Table 5). Although we expect models to benefit from higher quality samples and longer training sessions due to ShareGPT4V (Chen et al., 2023a), the results demonstrate that this is only consistently true when the model is finetuned with vision-text instruction data. More vision pretraining has a mixed effect on the text performance, boosting the 3B model's text score while hurting the 8B model's performance. Additionally, scaling up the model size effectively increases the text scores but leaves the vision scores roughly on the same level.

Diversity in Training Data In Section 2.1, we explored the similarity between instruction tuning using text-only and vision-language data. We now compare the mean cosine distances in two intra-dataset and one inter-datasets settings. Figure 4 reports the mean cosine similarities. The vision-language appears more homogeneous, with a higher mean, while the language data is more diverse. This observation aligns with the fact that vision-language datasets typically contain fewer distinct task types and tend to emphasize perceptual grounding, whereas language-only corpora encompass a broader spectrum. Importantly, the similarity scores between language-only and vision-language instructions are comparable to those within the language-only set, suggesting that diverse linguistic tasks inherently support better generalization—even across modalities. This could imply that language data, at least in our training data, better generalizes to vision datasets thanks to greater heterogeneity. Notably, though we use a diverse set of text-only and vision-language data, there is still a gap between the similarities, meaning text-only data that aligns better with vision-language can likely be constructed, which may improve performance even more.

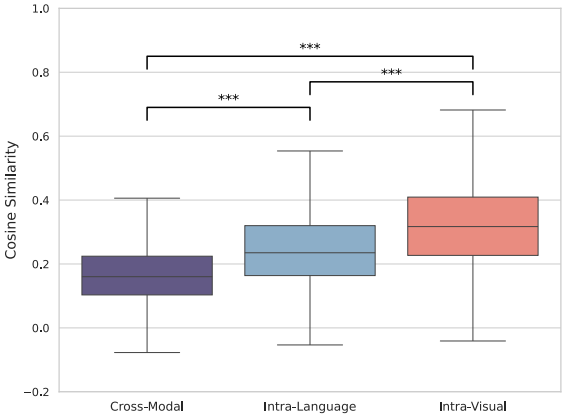


Figure 4: Distribution of the cosine similarity of random question pairs sampled in the language and vision-language settings. The stars (***) indicate significant differences ($p < 0.001$) between the mean similarity supported by the t-test.

4 Related Work

Multimodal large language models (MLLM) are language models endowed with the ability to use multiple modalities, such as images, videos, and audio (OpenAI, 2024; Meta AI, 2024; Team et al., 2023; OpenAI et al., 2024; Rubenstein et al., 2023; Zhang et al., 2023a; Ataallah et al., 2024; Bai et al., 2025; Li et al., 2024b; Liu et al., 2025b; Agrawal et al., 2024; Deitke et al., 2024; Chen et al., 2025). The most widely adopted are vision enhanced LLMs, where many design choices are

already extensively studied (Liu et al., 2023a; McKinzie et al., 2024; Lin et al., 2023; Laurençon et al., 2024; Tong et al., 2024; Karamcheti et al., 2024; Cocchi et al., 2025; Li et al.). A prevalent approach to building such MLLMs links pretrained visual encoders (Radford et al., 2021; Oquab et al., 2023) to LLMs (Touvron et al., 2023; Zheng et al., 2023; Chiang et al., 2023) via an adapter, thus transforming deep image features into soft prompts for the base LLM. In our work, we focus on one of the simplest yet high-performing and widely adopted MLLMs, using only a multi-layer perceptron as the adapter (Liu et al., 2023b,a, 2024a; Li et al., 2024a; Driess et al., 2023; Lin et al., 2023; Zeng et al., 2024).

Inspired by the success of instruction tuning in LLMs in zero-shot generalization (Wei et al., 2022; Wang et al., 2022a; Zhang et al., 2023c; Ouyang et al., 2022), following a pretraining step for vision-language feature alignment, there is a multimodal instruction tuning step to improve zero-shot performance on multimodal tasks (Xu et al., 2023; Li et al., 2024c). Notably, InstructBLIP (Dai et al., 2023) and LLaVA (Liu et al., 2023b) transform existing datasets into multimodal instructions using manual templates and synthetic data, a practice expanded upon in subsequent work (Tong et al., 2024; Chen et al., 2024b; Lin et al., 2023). Further work investigates how instruction tuning varies under different settings, e.g., how different components of the MLLM should learn differently during instruction tuning (Wu et al., 2024) and how instruction tuning works in a continual learning setting with many new tasks (Chen et al., 2024a). However, there lacks a comprehensive set of experiments that varies the composition of each modality in instruction tuning.

Though the primary goal of multimodal instruction tuning is to improve vision-language performance, text-only data is often included in both pretraining (McKinzie et al., 2024; Lin et al., 2023) and finetuning (Liu et al., 2023a; Huang et al., 2023; Bai et al., 2023; Ye et al., 2023, 2024; Luo et al., 2024; Lin et al., 2023; Tong et al., 2024; Dai et al., 2024; Bai et al., 2025; Li et al.; Zhang et al., 2024a,b) to prevent catastrophic forgetting and improve language performance. Many such papers disregard the impact of finetuning with text-only data on vision performance, focusing solely on language performance when ablating text-only data away, though there are notable exceptions (Huang et al., 2023; Ye et al., 2023, 2024; Lin et al., 2023; Dai et al., 2024; Zhang et al., 2024a). In these cases, there is modest evidence of transferability between modalities, where finetuning on both language and vision data exhibits about equal or better performance than training on one modality alone. However, in each of the existing work that finetune with text-only data alongside vision data, this performance boost is achieved by increasing the dataset size without consideration of how such data will increase the training cost (with the exception of Zhang et al. (2024a), which only tests with a low amount of text-only data). Hence, even though better performance is obtained when increasing the dataset size to train on text-only data, the instruction tuning step is more costly.

Due to the general cost of instruction tuning a MLLM, many approaches aims to decrease the cost of instruction tuning in the multimodal setting. These primarily include using lightweight adapters to decrease the number of parameters (Luo et al., 2024; Zhang et al., 2023b; Liu et al., 2025a) and choosing a subset of the training data using the MLLM itself or other methods (Chen et al., 2024c; Wei et al., 2023; Lee et al., 2024; Liu et al., 2024c; Safaei et al., 2025; Bi et al., 2025). A simpler way to decrease the cost is to instruction tune with a focus on text-only data. Since training on language instruction data is cheaper than training on the same number of vision instances, and language is foundational to the functioning of MLLMs, we focus on such a language-based approach.

5 Conclusion

We present MLAN, a language-based multimodal instruction tuning strategy for MLLMs that enhances zero-shot generalization and promotes effective knowledge transfer across modalities. We demonstrate—through controlled ablations under fixed training budgets—that language-based tuning establishes a robust knowledge foundation, even for tasks requiring visual understanding. Crucially, MLAN achieves strong performance on both language and vision benchmarks while significantly reducing reliance on image supervision. Our results show that language is not only sufficient but essential for efficient and generalizable multimodal learning. With MLAN, we hope to bring attention to the importance of language in MLLMs in visual instruction tuning, which we believe can be used in future work to improve training efficiency and performance.

6 Limitations

Our experiments are performed on models with the same multimodal architecture and pretraining procedure, not accounting for more advanced architecture or large-scale multimodal pretraining. Though we evaluate on a comprehensive set of vision-language benchmarks, we do not evaluate on specialized out of distribution tasks like OCR or captioning, focusing only on general tasks where the transferability is motivated. We invite future work to explore other methodologies to find where such specialized text-only and vision-language tasks align. Our analysis could also use experiments testing how instruction tuning varies when different tasks are trained on versus held-out, or on sequential finetuning versus sampling text-only and vision-language data. Furthermore, the instruction tuning experiments have the same data budget of 186,000 instances, while existing instruction tuning data may contain hundreds of thousands or even multi-million instances, which we leave to future work.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12B. *Preprint*, arXiv:2410.07073.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *Preprint*, arXiv:2404.03413.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *CoRR*, abs/2102.03315.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. PRISM: Self-Pruning Intrinsic Selection Method for Training-Free Multimodal Data Selection. *Preprint*, arXiv:2502.12119.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. 2024a. Coin: A benchmark of continual instruction tuning for multimodel large language model. *arXiv preprint arXiv:2403.08350*.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2024b. Visual instruction tuning with polite flamingo. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17745–17753. AAAI Press.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023a. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024c. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics.

Federico Cocchi, Nicholas Moratelli, Davide Caffagni, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2025. LLaVA-MORE: A Comparative Study of LLMs and Visual Backbones for Enhanced Visual Instruction Tuning. *Preprint*, arXiv:2503.15621.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *Preprint*, arXiv:2409.11402.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. *Preprint*, arXiv:2409.17146.

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, and 3 others. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2391–2401. Association for Computational Linguistics.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *CoRR*, abs/2405.02246.

Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995*.

Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: What else influences visual instruction tuning beyond data?

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. LLaVA-OneVision: Easy Visual Task Transfer. *Preprint*, arXiv:2408.03326.

Chen Li, Yixiao Ge, Dian Li, and Ying Shan. 2024c. Vision-language instruction tuning: A review and analysis. *Transactions on Machine Learning Research*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, and 8 others. Eagle 2: Building Post-Training Data Strategies from Scratch for Frontier Vision-Language Models. *Preprint*, arXiv:2501.14818.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. VILA: on pre-training for visual language models. *CoRR*, abs/2312.07533.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, and Cheng Han. 2025a. Re-Imagining Multimodal Instruction Tuning: A Representation View. *Preprint*, arXiv:2503.00723.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Yike Yuan, Wangbo Zhao, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision–ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pages 216–233. Springer.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, and 8 others. 2025b. NVILA: Efficient Frontier Visual Language Models. *Preprint*, arXiv:2412.04468.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024c. Less is more: Data value estimation for visual instruction tuning. *arXiv preprint arXiv:2403.09559*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2024. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, and 13 others. 2024. MM1: methods, analysis & insights from multimodal LLM pre-training. *CoRR*, abs/2403.09611.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

OpenAI. 2024. Hello gpt-4.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. Audiopalm: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. 2025. Filter Images First, Generate Instructions Later: Pre-Instruction Data Selection for Visual Instruction Tuning. *Preprint*, arXiv:2503.07591.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *CoRR*, abs/2406.16860.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, and 1 others. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.

Junda Wu, Xintong Li, Tong Yu, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Jingbo Shang, and Julian McAuley. 2024. Commit: Coordinated instruction tuning for multimodal large language models. *arXiv preprint arXiv:2407.20454*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 4818–4829. IEEE.

Zhiyang Xu, Chao Feng, Rulin Shao, Trevor Ashby, Ying Shen, dingnan jin, Yu Cheng, Qifan Wang, and Lifu Huang. 2024. Vision-flan: Scaling human-labeled tasks in visual instruction tuning. In *Annual Meeting of the Association for Computational Linguistics*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, 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, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. 2024. What matters in training a gpt4-style language model with multimodal inputs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7937–7964. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, and 4 others. 2024a. MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning. *Preprint*, arXiv:2409.20566.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023c. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qingguo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024b. Wings: Learning multimodal llms without text-only forgetting. *Advances in Neural Information Processing Systems*, 37:31828–31853.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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 Additional Implementation Details

For language-based instruction tuning, we use our carefully crafted dataset with tasks across modalities. To avoid data contamination, only the train split of each dataset is used for finetuning, and the test split, or the validation split if the test split is not publicly available, is reserved for evaluation. Similar to various multimodal instruction tuning work (Xu et al., 2023; Dai et al., 2023), we select unseen datasets of both modalities for evaluation. They are used to quantify performance in a general setting.

We maintain a fixed data budget of 186,000 instances throughout the training sessions. All training instances are sampled from Super-NaturalInstructions and Vision-Flan, according to the designated ratio. For the former, to prevent overfitting to a specific task, we sample an equal number of instances from every task. For the latter, since ScienceQA (Lu et al., 2022) is included in the training set, we manually remove them for evaluation purposes so there is no contamination. For finetuning, we apply the same chat template to all models in the following format: "USER:<query>ASSISTANT:<response>". The same prompt is used to format inputs during evaluation.

B Additional Training Details

We finetune pretrained MLLMs on the text-only data and denote those with a 75% text-only/25% vision-language split as MLAN. Acknowledging the recent trend of including a small portion of text-only data into vision instruction tuning data, we establish two additional baselines by finetuning on two separate versions of our training dataset that contain only 6% and 25% language instruction data, similar to the ratio in Liu et al. (2023a) and Tong et al. (2024). For a fair comparison, we limit the total number of training sequences in all settings to 186,000 samples from our training data.

C Dataset Summary

In Tables 6 and 7 we provide information about all 12 benchmarks used for evaluation. Note that in the main body we present results on 13 datasets, as we do not combine ARC-E and ARC-C.

D Additional Related Work

Our work focuses on choosing a simple multi-layer perception as the adapter in LLaVA (Liu et al., 2023b,a). In contrast, BLIP-2 (Li et al., 2023b) and Flamingo (Alayrac et al., 2022) design attention-based modules to attentively pool visual features, among a variety of other choices that combine existing methods or create new ones (Zhu et al., 2023; Chen et al., 2023b; Laurençon et al., 2024). To train the model, most often there is a pretraining step focusing on aligning the multimodal features with a modality connector (Yin et al., 2023), though some models are trained from scratch (Huang et al., 2023; Xiao et al., 2024). A main design choice in MLLMs is whether to freeze or unfreeze the LLM during finetuning. Unfreezing the LLM effectively prevents catastrophic forgetting by maintaining text-only performance (Meta AI, 2024; Driess et al., 2023; Alayrac et al., 2022), but results in worse vision-language performance (Lin et al., 2023; Dai et al., 2024). In our work, we show that with an unfrozen LLM, training on a strong language-based dataset on a fixed data budget improves performance across modalities. To evaluate MLLMs, there are a wide variety of vision-language tasks (Xu et al., 2023; Dai et al., 2023; Tong et al., 2024). However, Cambrian-1 (Tong et al., 2024) demonstrate that certain vision-language datasets, including some we used (AI2D and RealWorldQA), exhibit only a minor drop in performance of around 5% if vision is disabled, suggesting that current vision-language evaluations may be more language-focused. Though there is a need for more vision-centric analysis, this emphasizes how important language is in many vision tasks, a fact central to our work.

D.1 Text-Only Data in Existing Work

Table 8 lists dataset sizes as well as the splits between vision-language and text-only data in popular models that use both. We note that most models instruction tune with a majority of vision-language data, with the exception of Kosmos-1 (Huang et al., 2023) being a model that uses language alone, though it has an extensive pretraining step that differs from the simple MLLM adapter paradigm. Ultimately, many papers do not share their overall composition, and the ones that do vary greatly. We hope our work prompts the community to be more open in sharing their results and to do more work finding an effective and efficient ratio that can be used successfully across models.

Dataset	Modality	Split	Answer Type	Dataset Type	Size
ARC-Easy (Bhakthavatsalam et al., 2021)	Text	Test	Multiple Choice	Held-out	2.2k
ARC-Challenge (Bhakthavatsalam et al., 2021)	Text	Test	Multiple Choice	Held-out	1.2k
BoolQ (Clark et al., 2019)	Text	Validation	True/False	Held-out	3.2k
CommonsenseQA (Talmor et al., 2019)	Text	Validation	Multiple Choice	Held-out	9.7k
PIQA (Bisk et al., 2020)	Text	Validation	Multiple Choice	Held-out	16.1k
MMLU (Hendrycks et al., 2020)	Text	Test	Multiple Choice	Held-out	14.0k
RACE (Lai et al., 2017)	Text	Test	Multiple Choice	Held-out	1.05k
CosmosQA (Huang et al., 2019)	Text	Validation	Multiple Choice	Held-out	3.0k
POPE (Li et al., 2023c)	Vision	Test	True/False	Held-out	9.0k
ScienceQA-IMG (Lu et al., 2022)	Vision	Test	Multiple Choice	Held-out	5.0k
MMMU (Yue et al., 2024)	Vision	Validation	Multiple Choice	Held-out	1.5k
MME (Fu et al., 2023)	Vision	Test	True/False	Held-out	2.8k
MMBench (Liu et al., 2024b)	Vision	Dev	Multiple Choice	Held-out	5.2k

Table 6: Overview of evaluation datasets.

Dataset	Descriptions
CosmosQA (Huang et al., 2019)	Questions require reasoning based on people's everyday narratives to deduce the causes and effects of pertinent events.
CommonsenseQA (Talmor et al., 2019)	CommonsenseQA contains questions without context about understanding and relations between common objects.
ARC (Bhakthavatsalam et al., 2021)	ARC consists of grade-school level multiple-choice questions about understanding scientific concepts. Both easy and challenge splits are used.
RACE (Lai et al., 2017)	Race contains questions about long paragraphs collected from K12 English examinations in China.
BoolQ (Clark et al., 2019)	BoolQ asks whether a statement about a given long context is correct.
MMLU (Hendrycks et al., 2020)	A benchmark testing multi-task language understanding across 57 subjects, assessing model performance on expert-level multiple-choice questions.
PIQA (Bisk et al., 2020)	PIQA evaluates physical commonsense reasoning by selecting the most plausible solution to everyday scenarios.
MME (Fu et al., 2023)	MME is a multimodal benchmark for assessing cognition and perception capabilities of MLLMs across multiple domains with yes and no questions.
MMMU (Yue et al., 2024)	A multi-disciplinary benchmark testing on expert-level knowledge with vision and question queries. Questions types contain short response and multiple choice.
MMBench (Liu et al., 2024b)	A comprehensive multimodal benchmark that evaluates scientific knowledge with multiple choice questions.
POPE (Li et al., 2023c)	POPE asks to determine whether an object is present in the scene. We use adversarial, popular, and random splits for evaluation.
ScienceQA (Lu et al., 2022)	ScienceQA contains both vision-language and text-only questions about scientific concepts. We use all questions to test the overall ability of our models.

Table 7: Short descriptions for the evaluation benchmarks in our study.

Name	Text-Only Size	Total Size	text-only (%)
LLaVA-1.5 (Liu et al., 2023a)	40k	665k	6.0%
QwenVL (Bai et al., 2023)	N/A	350k	N/A
QwenVL2.5 (Bai et al., 2025)	∼1M	∼2M	50%
NVLM (Dai et al., 2024)	N/A	N/A	N/A
VILA (Lin et al., 2023)	1M	N/A	N/A
mPLUG-Owl (Ye et al., 2023)	242k	392k	61.7%
mPLUG-Owl2 (Ye et al., 2024)	558k	1.23M	45.4%
PrismaticVLM (Karamcheti et al., 2024)	40k	665k	6.0%
MM1 (McKinzie et al., 2024)	N/A	1.45M	N/A
MM1.5 (Zhang et al., 2024a)	–	–	10%
Kosmos-1 (Huang et al., 2023)	122.5k	122.5k	100%
LaVIN (Luo et al., 2024)	52k	204k	25.5%
Cambrian-1 (Tong et al., 2024) – Cambrian-7M	1.68M	∼7M	23.8%
Eagle 2 (Li et al.) – Stage 1.5	4.75M	21.6M	22.0%
LLaVA-OneVision (Li et al., 2024b) – Single-Image Data	457.6k	3.2M	14.3%

Table 8: Language instruction tuning dataset sizes in existing MLLMs. N/A means the number is either not presented in the paper or is unclear. A dash means the size is unclear.