Visual Instruction Tuning with Polite Flamingo

Delong Chen[1,2], **Jianfeng Liu**[1], **Wenliang Dai**[2], **Baoyuan Wang**[1]

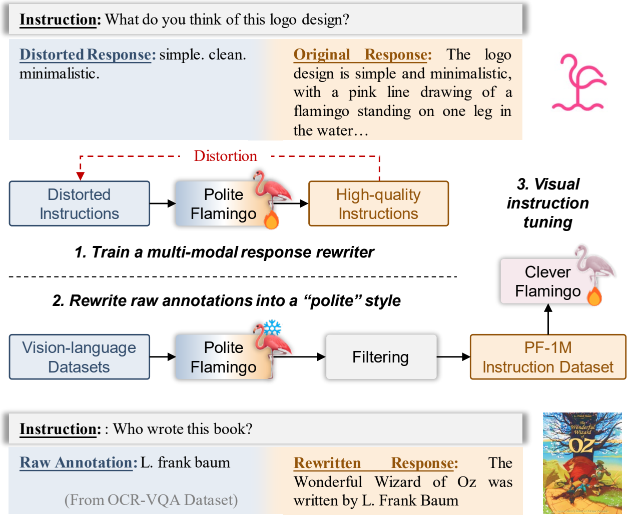
[1]Xiaobing.AI
[2]Centre for Artificial Intelligence Research (CAiRE),
Hong Kong University of Science and Technology
{delong.chen, wdaiai}@connect.ust.hk, {liujianfeng, wangbaoyuan}@xiaobing.ai

Abstract

Recent research has demonstrated that the multi-task fine-tuning of multi-modal Large Language Models (LLMs) using an assortment of annotated vision-language datasets significantly enhances their performance. Yet, during this process, a side effect, which we termed as the *"multi-modal alignment tax"*, surfaces. This side effect negatively impacts the model's ability to format responses appropriately - for instance, its "politeness" - due to the overly succinct and unformatted nature of raw annotations, resulting in reduced human preference. In this paper, we introduce Polite Flamingo, a multi-modal response rewriter that transforms raw annotations into a more appealing, "polite" format. Polite Flamingo is trained to reconstruct high-quality responses from their automatically distorted counterparts and is subsequently applied to a vast array of vision-language datasets for response rewriting. After rigorous filtering, we generate the PF-1M dataset and further validate its value by fine-tuning a multi-modal LLM with it. Combined with novel methodologies including U-shaped multi-stage tuning and multi-turn augmentation, the resulting model, Clever Flamingo, demonstrates its advantages in both multi-modal understanding and response politeness according to automated and human evaluations. Code and dataset are available at https://github.com/ChenDelong1999/polite-flamingo

Introduction

General-purpose AI systems have attracted a significant amount of interest due to their broad range of applications (*e.g.*, smart assistants). They are expected to be capable of accurately perceiving the visual world, comprehending diverse human requests, and providing helpful yet natural responses. Prior works towards this goal (*e.g,* OFA (Wang et al. 2022a), Unified-IO (Lu et al. 2022), Uni-Perceiver (Zhu et al. 2022)) have focused on training multi-modal transformers via multi-task learning, but they lack the generalization ability to unseen tasks or instructions, and they are not capable of offering user-friendly natural responses. Recently, instruction tuning empowers Large Language Models (LLMs) strong instruction-following and response formatting abilities, making it more convenient and efficient to access its encoded knowledge and complex reasoning ability. Many researchers attempted to connect visual representations with LLMs to transfer such powerful capability to vision-language tasks. Massive image-text data collected from the Internet can be used to train the visual representation (*e.g.,* CLIP (Radford et al. 2021)) and the connector (*e.g.,* Flamingo (Alayrac et al. 2022), Kosmos-1 (Huang et al. 2023), LLaVA (Liu et al. 2023b), MiniGPT-4 (Zhu et al. 2023)), but such supervision is usually noisy and could not cover much fine-grained information that encourages deeper visual understanding beyond shallow semantics. A promising direction is introducing annotated captioning / VQA / visual reasoning datasets, which exhibit a stronger alignment of real-world human needs than these image-text pairs sourced from the Internet. Concurrent works such as InstructBLIP (Dai et al. 2023), Otter (Li et al. 2023b), PaLI-X (Chen et al. 2023), and Ying-LM (Li et al. 2023c), have shown encouraging results of using a collection of vision-language datasets for visual instruction tuning.


Instruction: What do you think of this logo design?

Distorted Response: simple. clean. minimalistic.

Original Response: The logo design is simple and minimalistic, with a pink line drawing of a flamingo standing on one leg in the water…

Distortion
Distorted Instructions
Polite Flamingo
High-quality Instructions
3. Visual instruction tuning
1. Train a multi-modal response rewriter
Clever Flamingo
2. Rewrite raw annotations into a "polite" style
Vision-language Datasets
Polite Flamingo
Filtering
PF-1M Instruction Dataset

Instruction: : Who wrote this book?

Raw Annotation: L. frank baum

(From OCR-VQA Dataset)

Rewritten Response: The Wonderful Wizard of Oz was written by L. Frank Baum

Figure 1: Overview of our proposed approach.

However, there exists a significant challenge yet to be resolved in the process of visual instruction tuning. Existing captioning, VQA, and visual reasoning datasets typically provide very concise ground truths or answers. However, as human users, we generally prefer AI assistants that can provide ChatGPT-style structured responses, along with optional detailed explanations and elaborations. When using raw annotations for visual instruction tuning, their style would also be learned by the model, even the LLM part is kept frozen and only the connector is tuned. As a result, the InstructBLIP model, the current SoTA model on a wide range of vision-language benchmarks, ranked second to last (Li et al. 2023a) in Multi-Modality Arena (Xu et al. 2023), a user rating-based evaluation platform of multimodal LLMs. The model with the lowest Elo rating score is Multimodal-GPT (Gong et al. 2023), which is also tuned with raw annotations. This phenomenon is caused by the additional multi-modal alignment step upon LLM, which thus can be termed as "*multi-modal alignment tax*":

Definition 1. *Multi-modal alignment tax* $\Delta P_{\{g,f_{LLM}\}}$ *is the extra cost of enabling multi-modal perception for LLMs via visual instruction tuning* g *that maps a text-only* f_{LLM} *to a multimodal LLM, i.e.,* $g(f_{LLM}) \rightarrow f_{MLLM}$. *The cost is typically reflected as a degradation in task performance that measures model capacity from a certain perspective. Considering a total of* n *tasks* $\{T_1, T_2, ..., T_n\}$ *and their corresponding performance measure* P_{T_i}, *the multi-modal alignment tax can be quantified as* $\Delta P_{\{g,f_{LLM}\}} = \sum_{i=1}^{n} (P_{T_i}(f_{LLM}) - P_{T_i}(f_{MLLM}))$.

The root cause is that: visual representations are fed as soft prompts or prefixes to the LLM, while it is proved that prompt tuning or prefix tuning is able to drastically change the behavior of language models, similar to other parameter-efficient fine-tuning (PEFT) methods such as LoRA (Hu et al. 2022). In this paper, our goal is to prevent LLMs from learning undesired response styles of raw vision-language dataset annotations during visual instruction tuning, thus being a "*polite*" multi-modal LLM:

Definition 2. *Polite multi-modal LLMs provide natural and appropriate responses to user queries. Reduction in politeness is a specific instance of multi-modal alignment tax that impacts the model's ability to maintain optimal response styles.*

To achieve this goal, we introduce a novel method that involves converting these raw responses into natural ones, and we then train the multi-modal LLM using this style-transferred high-quality instruction data, thus mitigating the multi-modal alignment tax on response politeness. As shown in Figure 1, to obtain a rewriter that is capable of transferring the response style, we first distort the "polite" version of the response (*e.g.,* GPT-4 generated contents) into an "impolite" one, approximating the distribution of existing vision-language dataset annotations. We fine-tune a multi-modal LLM, OpenFlamingo-9B (Awadalla et al. 2023), to learn the reversed mapping (*i.e.,* impolite → polite). Subsequently, we apply the learned model, referred to as "Polite Flamingo", to rewrite massive annotations in existing vision-language datasets. After carefully filtering out low-quality results and hallucinations, we obtain a high-quality yet large-scale visual instruction tuning dataset PF-1M, and use it to tune a multi-modal LLM.

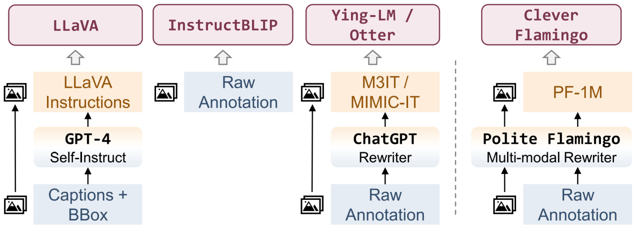


Figure 2: Comparison of different visual instruction tuning methods. LLaVA (Liu, Emerson, and Collier 2022) performs multi-modal self-instruct (Wang et al. 2022b) using GPT-4, which has high API cost and limited visual groundedness; InstructBLIP (Dai et al. 2023) directly uses learn raw annotations, and thus suffer from multi-modal alignment tax; M^{3}IT (Li et al. 2023c) and MIMIC-IT (Li et al. 2023a) employed ChatGPT-based rewriters, while we train a Polite Flamingo to rewrite responses, which enjoys advantages of 1) multi-modality, 2) scalability, and 3) diversity.

We perform a comprehensive evaluation comparing the resulting visual instruction-tuned model, which we called "Clever Flamingo", with other multi-modal LLMs, including MiniGPT-4 (Zhu et al. 2023), LLaVA (Liu, Emerson, and Collier 2022), InstructBLIP (Dai et al. 2023), and Otter (Li et al. 2023b). In summary, Clever Flamingo outperforms all of these models on detailed image captioning tasks, and only underperforms the InstructBLIP series (Dai et al. 2023) on VQA tasks (InstructBLIP uses a 3×heavier visual backbone, 8.6×larger pretraining dataset, and +0.6M more instruction samples). For multi-image reasoning tasks, Clever Flamingo outperforms the Otter baseline by a significant margin. In terms of human preference (*i.e.,* politeness), Clever Flamingo only underperforms the LLaVA series, which uses purely GPT-4-generated instructions. The contributions of this paper are summarized as follows:

- We proposed a novel method to curate raw vision-language datasets into visual instruction tuning data, which enables learning from a wide range of annotated datasets with reduced multi-modal alignment tax.
- We constructed a large-scale visual instruction tuning dataset based on response rewriting, and provide some empirical solutions to ensure data quality.
- We further introduced a U-shaped multi-stage visual instruction tuning pipeline and multi-turn augmentations to produce a strong multi-modal LLM efficiently.
- We performed comprehensive evaluations in terms of both multi-modal understanding and response politeness using automated evaluators, whose reliability is verified by human evaluations.

Related Works

Visual instruction tuning for multi-modal LLM. Research on enabling visual perception for powerful but blind LLMs attracted widespread attention recently (Yin et al. 2023a). The most straightforward methodology is to integrate image captioning experts via prompt engineering (*e.g.,* Socratic Models (Zeng et al. 2022), HuggingGPT (Shen et al. 2023), MM-REACT (Yang et al. 2023)) . However, this is inefficient due to the low bandwidth of natural language communication: given the diversity of real-world visual tasks, describing all of the potential task-relevant information within a single image requires a huge amount of language tokens. Therefore, many efforts opt to connect compact latent visual representations through a dense connector by visual instruction tuning, such as MiniGPT-4 (Zhu et al. 2023), LLaVA (Liu, Emerson, and Collier 2022), Multimodal-GPT (Gong et al. 2023), LLaMA-Adapter (Zhang et al. 2023), Otter (Li et al. 2023b), mPLUG-Owl (Ye et al. 2023), InstructBLIP (Dai et al. 2023). These models use linear projectors or perceivers as the connector between visual models and LLM, thus having a much larger information bandwidth compared to those prompt-based natural language communications.

Data for visual instruction tuning. However, what data is optimal for training these connectors to ensure that they propagate visual information faithfully is unclear. Existing attempts include generating self-instruct (Wang et al. 2022b) data (*i.e.,* LLaVA (Liu, Emerson, and Collier 2022)), using image-text captioning datasets (*e.g.,* COCO (Chen et al. 2015), SBU (Ordonez, Kulkarni, and Berg 2011), CC-3M (Sharma et al. 2018)), and unifying downstream vision-language datasets (*e.g.,* VQA and visual reasoning datasets). Although GPT-4 generated LLaVA dataset enjoy very high quality, its scale remains insufficient, and it could not encourage fine-grained vision-language alignment, as it does not "make V in VQA matter" (Goyal et al. 2017). On the other hand, using captioning datasets only would result in degenerated QA capabilities, as a soft prompt that encourages image captioning is implicitly learned by the connector, then the model would prefer to give an image caption even if the instruction asks it to answer a certain question.

Multi-modal alignment tax. Therefore, many efforts have been focused on utilizing downstream vision-language datasets, including Multimodal-GPT (Gong et al. 2023), Otter (Li et al. 2023b), InstructBLIP (Dai et al. 2023), M^{3}IT (Li et al. 2023c), LAMM (Yin et al. 2023b). Unfortunately, the multi-modal alignment tax (Definition 1) becomes a serious side effect that destroys the response formatting ability of the resulting multi-modal LLMs. To avoid such cost, the earliest work Multimodal-GPT (Gong et al. 2023) simply removed vision-language datasets that contain short answers. InstructBLIP (Dai et al. 2023) adds additional prompts such as "provide your answer as short as possible" to the instruction, but still could not mitigate the short answer bias due to the imbalance of response style – most responses in the training data are very short so the model just ignores these additional prompts.

ChatGPT-based text-only rewriter. Another attempt to mitigate the multi-modal alignment tax is to use ChatGPT to rewrite the short answer, as adopted in concurrent works M^{3}IT (Li et al. 2023c) and MIMIC-IT (Li et al. 2023a). We compare our method with them in Figure 2. Since our Polite Flamingo is a *multi-modal* rewriter, it can fuse visual perception with text semantics to rewrite, as opposed to these ChatGPT-based blind models that can only rely on the answer information. Polite Flamingo is also much lighter, cheaper, and does not require any API cost, leading to better scalability[1]. Moreover, Polite Flamingo is specially trained on 255k diverse rewriting examples, while ChatGPT can only perform zero-shot or few-shot rewriting. As an example of its limitation, M^{3}IT (Li et al. 2023c) used a single in-context rewriting demonstration to prompt ChatGPT, which resulted in limited diversity – 96% rewritten samples within its A-OKVQA subset have the sentence pattern of "`{rational}, so the answer is {answer}`". Finally, our work also shares some similarities with FuseCap (Rotstein et al. 2023) and LaCLIP (Fan et al. 2023) and RemoteCLIP (Liu et al. 2023a) that generate/rewrite image captions to train vision language models.

Polite Flamingo: a Multi-modal Instruction Response Rewriter

To learn a rewriter for raw annotations of vision-language datasets, the most straightforward way could be to train a model to directly predict a "polite" version from the corresponding raw annotations. Unfortunately, careful annotation of such translations is highly expensive and hard to scale. To overcome this limitation, we design a surrogate task that trains the rewriter to learn the style from existing high-quality instruction data, such as the LLaVA self-instruct dataset (Liu et al. 2023b). Specifically, we first transfer the style of these high-quality responses into low-quality ones, approximating the distribution of the raw annotations in the vision-language dataset that needs to be rewritten. Then, we train the model to reconstruct the original high-quality response from given distortions, as shown in Figure 3.

Our methodology is inspired by denoising AutoEncoder-style image enhancement models. These systems automatically introduce distortions, such as random noise or down-sampling, to the original images, and then the model is trained to reconstruct the original images. The resulting model can then be applied to image denoising or super-resolution. The key assumption of these image enhancement models, as well as our Polite Flamingo is that the distortion module should produce samples *i.i.d.* to the input samples during inference (*i.e.,* noise/low-resolution images, or raw annotations) so that the train-test domain divergence is small and these denoising AutoEncoders can generalize well.

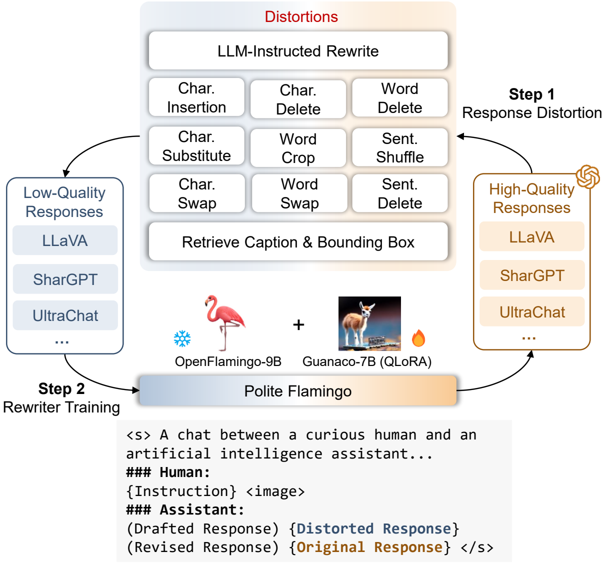
Response Distortion

To approximate the distribution of raw vision-language dataset annotations that would be used for Polite Flamingo inference, we develop the following three strategies for response distortion. Resulting examples can be found in the Appendix[2].

- **LLM-instructed Distortion**. Representative patterns of raw annotations include short answers (*e.g.,* VQA-

[1] Polite Flamingo can be run on consumer GPUs: BF-16 inference roughly takes 18 GB GPU memory.

[2] The appendix can be found at https://arxiv.org/pdf/2307.01003.pdf



```
<s> A chat between a curious human and an
artificial intelligence assistant...
### Human:
{Instruction} <image>
### Assistant:
(Drafted Response) {Distorted Response}
(Revised Response) {Original Response} </s>
```

Figure 3: Training pipeline of Polite Flamingo. We distort original high-quality responses into the corresponding low-quality version, then train a multi-modal LLM to predict the original response. This model is then used to rewrite raw annotations of a wide range of vision-language datasets and derive a PF-1M dataset for visual instruction tuning.

v2 (Goyal et al. 2017)), lacking punctuation or capitalization (*e.g.,* MS-COCO Captions (Chen et al. 2015)), not being coherent (*e.g.,* A-OKVQA (Schwenk et al. 2022)), etc., and we prompt an LLM to produce responses similar to these patterns. For each sample, we append another round of conversation, asking the model to transfer the original response into a "impolite" one. Furthermore, we randomly sample a distortion command from a pool containing a total of 24 alternatives and add it to the prompt with a probability of 50

- **Random Text Augmentations**. This distortion is much cheaper compared to LLM-based distortion, and we introduce it to further increase the diversity of the Polite Flamingo training set. Specifically, We use the `NLPAUG` library to perform character-level, word-level, and sentence-level text augmentation. Every level of augmentation is applied with a probability of 50
- **Retrieve Captions & Bounding Boxes**. In the LLaVA dataset (Liu et al. 2023b), GPT-4 is used to produce high-quality detailed captions for visual instruction tuning, given five captions and all bounding box annotations of each image. However, possibly due to the high API cost, there are only 23k samples of such detailed descriptions. Here we would like to distill such capability into the Polite Flamingo, and extrapolate it into the remaining MS-COCO samples, as well as other datasets with multiple captions (*e.g.,* Flicker-30k) or bounding box annotations (detection datasets). We retrieve the original captions and object bounding boxes in the `LLaVA-detailed-23k` dataset and use them as the distorted version with respect to the original detailed descriptions. We also insert the description of "The followings are specific object locations..." which was used for prompting GPT-4, to help Polite Flamingo understand bounding box annotations.

Training a Rewritter

We gathered a total of 255k samples to train the Polite Flamingo (see Appendix for details). We initialize the model from OpenFlamingo-9B (Awadalla et al. 2023), and insert a LoRA (Hu et al. 2022) adapter (initialized from the QLoRA of Guanaco-7B (Dettmers et al. 2023)) into its LLaMA-7B (Touvron et al. 2023) language model. We tune the LoRA weights only, and keep other parameters (*i.e.,* language model, ViT, perceiver, X-ATTN layers (Alayrac et al. 2022)) frozen to prevent overfitting. As shown in Figure 3, we provide the instruction, image, and distorted response to the Polite Flamingo, and ask it to predict the original response. Language modeling loss is only applied to the tokens corresponding to the original response.

Scale Up Visual Instruction Tuning with Polite Flamingo

Source Datasets

To scale up the vision-language instruction tuning data thus improving the visual understanding capability of the multi-modal LLM, we leverage the trained Polite Flamingo to rewrite the raw annotations of numerous vision-language datasets into polite responses. Similar to several concurrent works (Dai et al. 2023; Li et al. 2023c,a), we standardize them into a unified instruction-response format. The adopted datasets can be roughly divided into two main groups: captioning datasets, which task the model with providing detailed descriptions of image content, and VQA datasets, which require the model to accurately answer specific queries. We adopted a total of 37 datasets, see the appendix for a detailed summarization.

Filtering Strategies

Our rewriter, Polite Flamingo, is based on LLaMA-7B (Touvron et al. 2023), which is a relatively small language model. Through empirical observation, we have identified that Polite Flamingo is not a flawless response rewriter. It occasionally leaves the answer unchanged, produces repetitive patterns, or even changes the original answer and introduces hallucinated content. We design an automatic filtering pipeline to mitigate these problems and guarantee the quality of visual instruction tuning data. We use several rule-based filters, and several newly introduced model-based filters to measure the semantics of rewritten response, including a Semantic Textual Similarity (STS) model-based filter, a Natural Language Inference (NLI) model-based filter, and a CLIPScore-based hallucination filter. Please see Appendix for more details.

U-shaped Multi-stage Visual Instruction Tuning

We first leverage the Polite Flamingo to rewrite the response of source datasets (Section), obtaining 1.17M sam-

Method	#Instruction	Visual (#Params)	Connector (#Samples)	LLM (#Params)	Detailed Image Description			Visual Question Answering		
					COCO	TextCaps	Img2P	OK-VQA	VSR	Grid-3D
MiniGPT-4	3.5k	ViT-g (1.0B)	Linear (5M)	7B	14.4	15.5	14.7	10.4	14.0	19.0
				13B	23.1	19.2	23.7	23.8	24.6	20.0
LLaVA	177k	ViT-L (0.3B)	Linear (595k)	7B	23.8	21.1	23.6	32.1	36.1	20.8
				13B	23.1	20.7	23.2	30.9	34.1	22.5
InstructBLIP (Vicuna)	1.6M	ViT-g (1.0B)	BLIP-2 (129M)	7B	23.7	22.2	22.2	51.5	48.5	28.9
				13B	23.5	19.7	22.1	**52.2**	**48.9**	27.5
Otter	2.8M	ViT-L (0.3B)	OF-9B (15M)	7B	22.6	19.7	22.4	28.7	28.7	13.5
Ours	1.0M			7B	**24.3**	**24.1**	**24.7**	43.3	43.6	**29.0**
±Δ	**-1.8M**	-	-	-	**+1.7**	**+4.4**	**+2.3**	**+14.6**	**+14.9**	**+15.5**

Table 1: Performance comparison of with different multi-modal LLMs. We use Rouge-L as the metric for detailed image description tasks, and we use an NLI-based evaluator for VQA datasets. Blue numbers are results on unseen datasets (*i.e.,* zero-shot), and black numbers are results on unseen samples (*i.e.,* validation split of datasets seen during training). The bottom row (±Δ) compares our Clever Flamingo with Otter, which uses the same OF-9B (OpenFlamingo) as the base model.

ples. After filtering, 0.97M samples remained, which we refer to as the PF-1M dataset. In addition to PF-1M, we also adopt several high-quality text-only instruction datasets, since our base model OpenFlamingo-9B is based on the vanilla LLaMA-7B which is not instruction-tuned. Recent studies have shown that data quality is of vital importance during instruction tuning. Motivated by this, we consider the following datasets: UltraChat (Ding et al. 2023), ShareGPT, OASST-1 (Köpf et al. 2023), Alpaca-GPT-4 (Peng et al. 2023), GPTeacher, and InstructionWild (Xue et al. 2023). Together with PF-1M and LLaVA-instruction-177k, we have a total of 1.5M instruction data.

However, the samples in this dataset collection provide benefits to the model from very different perspectives. Text-only instructions enable the model to comprehend human requests and generate helpful responses in a proper style, while PF-1M data primarily facilitate the model in improving precise visual perception. To enhance training efficiency, we propose a U-shaped visual instruction tuning approach that encompasses three stages:

Stage 1 focuses on improving the instruction-following ability of the model by tuning only the language model (with LoRA). We utilize a total of 0.77M samples, which include all text-only instructions, LLaVA instructions, and 10% samples (97k) from PF-1M, and trained the model for a single epoch. The model is trained with a large context window of 1024 tokens. **Stage 2** shifts to improving the visual understanding capability of the model. We freeze the LoRA adapter and exclusively tune the connector using the entire PF-1M dataset. To enhance training efficiency, we use a smaller context window of 196 tokens. **Stage 3** uses the same setting as Stage 1, but we adjust the learning rate to 10× lower. The objective of Stage 3 is to fine-tune the model to recover the optimal politeness of the responses. This adjustment is necessary as the PF-1M dataset used in Stage 2 is generated by a 7B language model, which has lower quality than larger LLM-generated text-only instructions.

Multi-turn Augmentation

Given the diversity of instruction data, the length of each sample varies a lot. When using a large context window, short instruction samples would append many `<PAD>` tokens and waste a lot of computation. To address this, we introduce multi-turn augmentation, which involves randomly selecting instruction samples and concatenating them to form a multi-turn conversation. In this augmentation scheme, only the tokens corresponding to the response in each turn are considered when calculating the language modeling loss. This multi-turn also encourages the model to attend to the correct image for multi-turn multi-image conversations.

Evaluations

Evaluation of PF-1M Dataset

We analyze the improvement of "politeness" of Polite Flamingo rewriting (from raw annotations to PF-1M) through a quantitative evaluation. We assume that a reward model which is trained on human-labeled user preference data is able to provide an estimation of politeness. Results[3] show that Polite Flamingo significantly boosts the politeness of raw dataset annotations (from -2.42 to -0.50), and the resulting PF-1M outperforms the recently proposed large-scale instruction tuning dataset M^{3}IT (Li et al. 2023c) by a notable margin. Unfortunately, PF-1M cannot match those datasets produced by much larger LLM, especially those generated by GPT-4 (*i.e.,* LLaVA (Liu et al. 2023b) and Alpaca-GPT-4 (Peng et al. 2023)). But on the other hand, PF-1M is approximately 6× larger than the LLaVA dataset, and many LLaVA instructions are QA conversations under the theme of the image. In comparison, the PF-1M dataset is derived from annotated vision-language dataset and involves challenging samples that encourage fine-grained visual understanding. In addition, we also provide a qualitative evaluation of Polite Flamingo's rewriting in the Appendix.

Performance Comparison

We verify the performance of the Clever Flamingo by comparing it with other existing multi-modal LLMs. We focus on answering the following questions: 1) how well does it perform on vision-language tasks, 2) how does it generalize to unseen datasets, and 3) whether it produces human-

[3] The visualization of reward score distribution can be found at https://arxiv.org/pdf/2307.01003.pdf

Model	Spot-the-Diff		Image-editing		NLVR2	
	STS	Rouge	STS	Rouge	STS	Rouge
L.B.	31.6	0.119	13.9	0.023	7.0	0.012
Otter	39.5	0.129	33.2	0.136	11.5	0.069
Ours	**46.1**	**0.185**	**37.0**	**0.156**	**28.2**	**0.155**
±Δ	**+6.6**	**+.057**	**+3.9**	**+.020**	**+16.7**	**+.085**

Table 2: Multi-image reasoning tasks. "STS" means semantic textual similarity. The lower bound performance (L.B.) comes from a single-image model (InstructBLIP). Blue numbers indicates unseen datasets and black numbers correspond to results on unseen samples (*i.e.,* validation split).

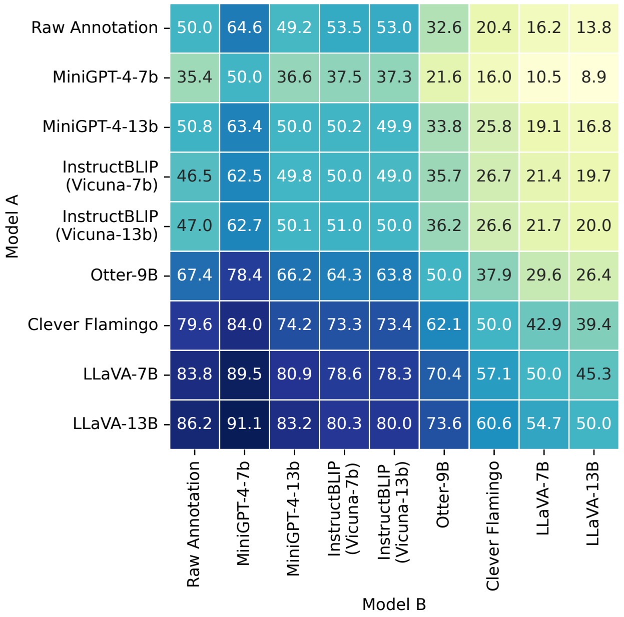


Figure 4: Win rate matrix of model A beat model B in terms of reward model score. For example, Clever Flamingo has a 62.1% win rate against Otter. Our model has a >50% win rate against other multi-modal LLMs despite the LLaVA series, which is trained on purely GPT-4 generated data.

preferred responses (*i.e.,* being polite). We first compare it with other models on image captioning and VQA tasks, then we present the evaluation of multi-image reasoning tasks, and finally, we analyze the politeness of these multi-modal LLMs.

Image Captioning and VQA Table 1 summarized the evaluation results comparing Clever Flamingo with other multi-modal LLMs on detailed image captioning and visual question answering . We use Rouge-L as the metric for captioning datasets and use an NLI model-based automated evaluator for VQA datasets (see appendix for more details). As our work is concurrent with InstructBLIP (Dai et al. 2023) and Otter (Li et al. 2023b), the dataset splitting (*i.e.,* assignments of held-in training datasets and held-out unseen testing datasets) is not fully aligned. We marked the held-in datasets with **black** and marked the held-out datasets with **blue**.

In summary, Clever Flamingo outperforms other counterparts on all three detailed image description datasets and the Grid-3D dataset, and only underperforms the InstructBLIP series on OK-VQA and VSR. Importantly, the settings (*e.g.,* the base model and training data amount) of these comparisons are not aligned. For InstructBLIP, a BERT-based Q-Former is firstly trained with BILP-generated and filtered 129M samples for two stages (about 3-4 epochs), then the model is instruction-tuned on a 1.6M collection of downstream data. In comparison, our Clever Flamingo, as well as the Otter model, is tuned from OpenFlamingo-9B, which uses a 3×smaller visual encoder, a lighter perceiver as the connector, and much less pre-training image-text data (15M) and training steps (single epoch) . When come to a fair comparison between Clever Flamingo and Otter (despite instruction data, Clever Flamingo uses 1.8M fewer data), our model outperforms Otter on every dataset, both held-in and held-out, by a significant margin.

Multi-image Reasoning Now we analyze the performance on multi-image reasoning tasks. We compare Clever Flamingo with Otter (Li et al. 2023b), which is also tuned from OpenFlamingo-9B – the only currently publicly available base multi-modal LLM that can process interleaved image-text data. The following three datasets are used for evaluation: 1) Spot-the-diff (Jhamtani and Berg-Kirkpatrick 2018), a change captioning dataset for surveillance camera imagery, 2) Image-editing-requests (Tan et al. 2019), which requires the model to infer image editing requests (*e.g,* Photoshop editing) given image pairs, and 3) Natural Language Visual Reasoning-2 (NVLR2) (Suhr et al. 2019), where the model needs to reason whether a statement holds true given two images.

We use Rouge-L between model prediction and ground truth as the metric. We further introduced a model-based evaluator "STS" (semantic textual similarity), which is measured by the cosine distance of sentence features , to compare high-level semantics of answer and ground truth (Reimers and Gurevych 2019). We also provide the evaluation result of a single-image model (InstructBLIP) as the lower bound. The result is shown in Table 2. Again, Clever Flamingo outperforms Otter on all three datasets by a large margin.

Politeness We used a reward model to evaluate the politeness of model responses on a total of 52k samples sourced from the validation/test split of a collection of vision-language downstream datasets . For each sample, we first obtain the prediction of multi-modal LLMs, then feed the instruction and the generated responses to a reward model to get reward scores, and make a pairwise comparison of the scores. In Figure 4, we visualize the average win rate – the statics of the pairwise comparison of all 52k samples. We also calculate the reward score of raw annotations for comparison.

As it can be seen, our Clever Flamingo is more likely to be preferred by the reward model (having >50% win rate) compared to all of the other compared multi-modal LLMs, except the LLaVA series. This is a direct result of the differ-

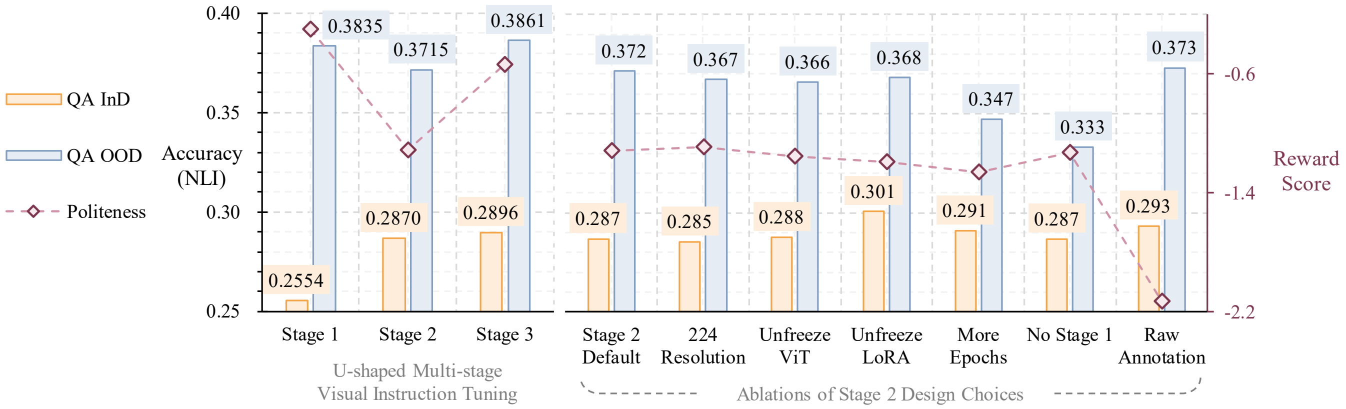


Figure 5: Results of ablation experiments on U-shaped multi-stage visual instruction tuning (left) and design choices in stage 2 (right). We calculate averaged NLI-based accuracy for held-in QA datasets (QA InD) and held-out datasets (QA OOD). We also report the average reward score to reflect the politeness of each alternative.

ences in instruction data: GPT-4 generated LLaVA dataset outperforms the PF-1M dataset in terms of reward score.

Ablation Study

We now present the ablation experiments to verify the effectiveness of various design choices of Clever Flamingo. We report the averaged NLI-based validation accuracy of in-domain (held-in) VQA datasets and out-of-distribution (held-out) VQA datasets, and further calculate the averaged reward score as a measurement of politeness.

The results are shown in Figure 5. On the left side, we first visualize the change of metrics during the U-shaped multi-stage visual instruction tuning. It shows that stage 2 boosts the in-domain QA accuracy, but also results in a degenerated politeness. Stage 3 maintains the in-domain QA accuracy, but recovers the politeness significantly. It is interesting to observe that OOD QA accuracy also exhibits a U-shaped curve. It seems that stage 2 led to sight overfitting to the PF-1M data distribution, well stage 3 alleviates this problem.

The right side of Figure 5 shows ablations on the Clever Flamingo stage 2. The observations on different alternatives are listed as follows. **1) 224 Resolution**: changing image resolution from default 336×336 to 224×224 hurt the performance, confirmed the hypothesize in (Liu et al. 2023c). **2) Unfreeze ViT**: further tuning ViT in addition to perceiver and XATTN failed to improve the performance significantly, and resulted in slight overfitting. It shows that the scale of PF-1M is still insufficient to support continual representation learning of the visual backbone. **3) Unfreeze LoRA**: this ablation significantly improved the PF-1M in-domain accuracy, but also hurt the generalization ability. **4) More Epochs**: we doubled the stage 2 epochs from 3 to 6, and found that it significantly hurt the generalization ability to the unseen domain. **5) No Stage 1**: when skipping stage 1 and directly going into stage 2 from vanilla OpenFlamingo-9B, the OOD generalization ability further dropped. It demonstrates that instruction samples used in stage 1 and stage 3 can effectively boost/maintain the OOD generalization ability. **6) Raw Annotation**: when skipping the Polite Flamingo-based rewriting and using the raw annotations in PF-1M for visual instruction tuning, both held-in and held-out accuracy got slightly improved, however, the multi-modal alignment tax is significant – the "politeness" dropped significantly.

Conclusion

This paper presents our solution to the multi-modal alignment tax problem, specifically, we want to use a diverse collection of downstream vision-language datasets to improve the visual understanding capability of multi-modal LLMs while avoiding the unformatted raw annotations to decrease the "politeness" of model responses. We trained a rewriter and used it to build a large-scale visual instruction tuning dataset. Incorporating U-shaped multi-stage tuning and multi-turn augmentation, we derived a strong multi-modal LLM , which has advantages in terms of both multi-modal understanding and response politeness.

References

Alayrac, J.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J. L.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo.

Chen, X.; Djolonga, J.; Padlewski, P.; Mustafa, B.; Changpinyo, S.; Wu, J.; Ruiz, C. R.; Goodman, S.; Wang, X.; Tay, Y.; Shakeri, S.; Dehghani, M.; Salz, D.; Lucic, M.; Tschannen, M.; Nagrani, A.; Hu, H.; Joshi, M.; Pang, B.; Montgomery, C.; Pietrzyk, P.; Ritter, M.; Piergiovanni, A. J.; Minderer, M.; Pavetic, F.; Waters, A.; Li, G.; Alabdulmohsin, I.;

Beyer, L.; Amelot, J.; Lee, K.; Steiner, A. P.; Li, Y.; Keysers, D.; Arnab, A.; Xu, Y.; Rong, K.; Kolesnikov, A.; Seyedhosseini, M.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2023. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *CoRR*, abs/2305.18565.

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR*, abs/1504.00325.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. C. H. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR*, abs/2305.06500.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *CoRR*, abs/2305.14314.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. *CoRR*, abs/2305.14233.

Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023. Improving CLIP Training with Language Rewrites. *CoRR*, abs/2305.20088.

Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; and Chen, K. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *CoRR*, abs/2305.04790.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6325–6334. IEEE Computer Society.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O. K.; Patra, B.; Liu, Q.; Aggarwal, K.; Chi, Z.; Bjorck, J.; Chaudhary, V.; Som, S.; Song, X.; and Wei, F. 2023. Language Is Not All You Need: Aligning Perception with Language Models. *CoRR*, abs/2302.14045.

Jhamtani, H.; and Berg-Kirkpatrick, T. 2018. Learning to Describe Differences Between Pairs of Similar Images. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, 4024–4034. Association for Computational Linguistics.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z.; Stevens, K.; Barhoum, A.; Duc, N. M.; Stanley, O.; Nagyfi, R.; ES, S.; Suri, S.; Glushkov, D.; Dantuluri, A.; Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant Conversations - Democratizing Large Language Model Alignment. *CoRR*, abs/2304.07327.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Yang, J.; Li, C.; and Liu, Z. 2023a. MIMIC-IT: Multi-Modal In-Context Instruction Tuning. *CoRR*, abs/2306.05425.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; and Liu, Z. 2023b. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *CoRR*, abs/2305.03726.

Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; Kong, L.; and Liu, Q. 2023c. M^{3}IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *CoRR*, abs/2306.04387.

Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; and Zhou, J. 2023a. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *CoRR*, abs/2306.11029.

Liu, F.; Emerson, G.; and Collier, N. 2022. Visual Spatial Reasoning. *CoRR*, abs/2205.00363.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. *CoRR*, abs/2304.08485.

Liu, Y.; Li, Z.; Li, H.; Yu, W.; Huang, M.; Peng, D.; Liu, M.; Chen, M.; Li, C.; Jin, L.; and Bai, X. 2023c. On the Hidden Mystery of OCR in Large Multimodal Models. *CoRR*, abs/2305.07895.

Lu, J.; Clark, C.; Zellers, R.; Mottaghi, R.; and Kembhavi, A. 2022. Unified-IO: A Unified Model for Vision, Language, and Multi-Modal Tasks. *CoRR*, abs/2206.08916.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F. C. N.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, 1143–1151.

Peng, B.; Li, C.; He, P.; Galley, M.; and Gao, J. 2023. Instruction Tuning with GPT-4. *CoRR*, abs/2304.03277.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *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*, 8748–8763. PMLR.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rotstein, N.; Bensaïd, D.; Brody, S.; Ganz, R.; and Kimmel, R. 2023. FuseCap: Leveraging Large Language Models to Fuse Visual Data into Enriched Image Captions. *CoRR*, abs/2305.17718.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge. In Avidan, S.; Brostow, G. J.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision - ECCV 2022 - 17th European*

Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII, volume 13668 of *Lecture Notes in Computer Science*, 146–162. Springer.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2556–2565. Association for Computational Linguistics.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. *CoRR*, abs/2303.17580.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 6418–6428. Association for Computational Linguistics.

Tan, H.; Dernoncourt, F.; Lin, Z.; Bui, T.; and Bansal, M. 2019. Expressing Visual Relationships via Language. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, 1873–1883. Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022a. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 23318–23340. PMLR.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2022b. Self-Instruct: Aligning Language Model with Self Generated Instructions. *CoRR*, abs/2212.10560.

Xu, P.; Shao, W.; Zhang, K.; Gao, P.; Liu, S.; Lei, M.; Meng, F.; Huang, S.; Qiao, Y.; and Luo, P. 2023. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *CoRR*, abs/2306.09265.

Xue, F.; Jain, K.; Shah, M. H.; Zheng, Z.; and You, Y. 2023. Instruction in the Wild: A User-based Instruction Dataset. https://github.com/XueFuzhao/InstructionWild.

Yang, Z.; Li, L.; Wang, J.; Lin, K.; Azarnasab, E.; Ahmed, F.; Liu, Z.; Liu, C.; Zeng, M.; and Wang, L. 2023. MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. *CoRR*, abs/2303.11381.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qi, Q.; Zhang, J.; and Huang, F. 2023. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *CoRR*, abs/2304.14178.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023a. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

Yin, Z.; Wang, J.; Cao, J.; Shi, Z.; Liu, D.; Li, M.; Sheng, L.; Bai, L.; Huang, X.; Wang, Z.; Shao, J.; and Ouyang, W. 2023b. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. *CoRR*, abs/2306.06687.

Zeng, A.; Wong, A.; Welker, S.; Choromanski, K.; Tombari, F.; Purohit, A.; Ryoo, M. S.; Sindhwani, V.; Lee, J.; Vanhoucke, V.; and Florence, P. 2022. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. *CoRR*, abs/2204.00598.

Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; and Qiao, Y. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *CoRR*, abs/2303.16199.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *CoRR*, abs/2304.10592.

Zhu, X.; Zhu, J.; Li, H.; Wu, X.; Li, H.; Wang, X.; and Dai, J. 2022. Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 16783–16794. IEEE.