

Visual Instruction Tuning with Polite Flamingo

Delong Chen¹ Jianfeng Liu¹ Wenliang Dai² Baoyuan Wang¹
¹Xiaobing.AI ²Hong Kong University of Science and Technology

Abstract

Recent research has demonstrated that the multi-task fine-tuning of multi-modal Large Language Models (LLMs) using an assortment of annotated downstream 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.¹

1. 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 [1], Unified-IO [2], Uni-Perceiver [3]) 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 [4] empowers Large Language Models (LLMs) [5] strong instruction-

¹https://github.com/ChenDelong1999/polite_flamingo

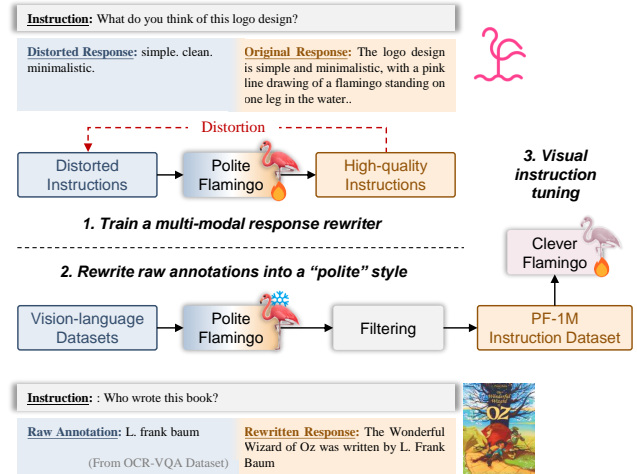


Figure 1. **Overview of our approach.** We first train a multi-modal response rewriter on high-quality instruction datasets, resulting in a "Polite Flamingo" capable of translating raw annotations in vision-language datasets into a "polite" style. After careful filtering, we use the rewritten data PF-1M for visual instruction tuning, and obtained a strong multi-modal LLM called *Clever Flamingo*.

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 [6]) and the connector (e.g., Flamingo [7], Kosmos-1 [8], LLaVA [9], MiniGPT-4 [10]), 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 captions sourced from the Internet. Concurrent works such as InstructBLIP [11], Otter [12], PaLI-X [13], and Ying-LM [14], have shown encouraging results of using a collection of vision-language datasets for visual instruction tuning.

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 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 [15] in Multi-Modality Arena [16], a user rating-based evaluation platform of multi-modal LLMs. The model with the lowest Elo rating score is Multimodal-GPT [17], 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 1.1. *Multi-modal alignment tax is the extra cost of enabling or improving multi-modal perception for LLMs. The cost is typically reflected as a degradation in performance from certain perspectives.*

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 [18], similar to other parameter-efficient fine-tuning (PEFT) methods such as LoRA [19]. 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 1.2. *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 [20], to learn the reversed mapping (i.e., impolite \rightarrow 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.

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 [10], LLaVA [21], InstructBLIP [11], and Otter [12]. In summary, Clever Flamingo outperforms all of these models on detailed image captioning tasks, and only underperforms the InstructBLIP series [11] on VQA tasks (InstructBLIP uses a 3 \times heavier visual backbone, 8.6 \times 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 [21], 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 empirical solutions to ensure data quality and mitigate hallucinations.
- We further introduced a U-shaped multi-stage visual instruction tuning pipeline and multi-turn augmentations to produce a strong instruction-tuned 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.

2. Related Works

Visual instruction tuning for multi-modal LLM. Research on enabling visual perception for powerful but blind LLMs attracted widespread attention recently [23]. The most straightforward methodology is to integrate image captioning experts via prompt engineering (e.g., Socratic Models [24], HuggingGPT [25], MM-REACT [26]). 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

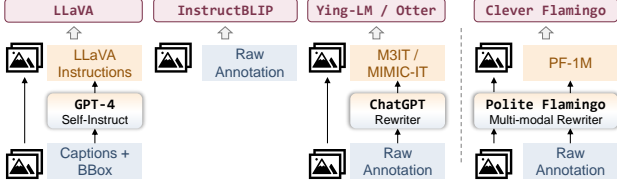


Figure 2. **Comparison of different visual instruction tuning methods.** LLaVA [21] performs multi-modal self-instruct [22] using GPT-4, which has high API cost and limited visual groundness; InstructBLIP [11] directly uses learn raw annotations, and thus suffer from multi-modal alignment tax; M³IT [14] and MIMIC-IT [15] 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.

compact latent visual representations through a dense connector by visual instruction tuning, such as MiniGPT-4 [10], LLaVA [21], Multimodal-GPT [17], LLaMA-Adapter [27], Otter [12], mPLUG-Owl [28], InstructBLIP [11]. 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 [22] data (*i.e.*, LLaVA [21]), using image-text captioning datasets (*e.g.*, COCO [29], SBU [30], CC-3M [31]), 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” [32]. 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 [17], Otter [12], InstructBLIP [11], M³IT [14], LAMM [33]. Unfortunately, the multi-modal alignment tax (Definition 1.1) 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 [17] simply removed vision-language datasets that contain short answers. InstructBLIP [11] 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 M³IT [14] and MIMIC-IT [15]. 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 scalability². 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, M³IT [14] 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 [34] and LaCLIP [35] and RemoteCILP [36] that generate/rewrite image captions to train vision language models.

3. 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 [9]. 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,

²Polite Flamingo is based on LLaMA-7B and can be run on consumer GPUs. BF-16 inference of Polite Flamingo roughly takes 18 GB GPU memory.

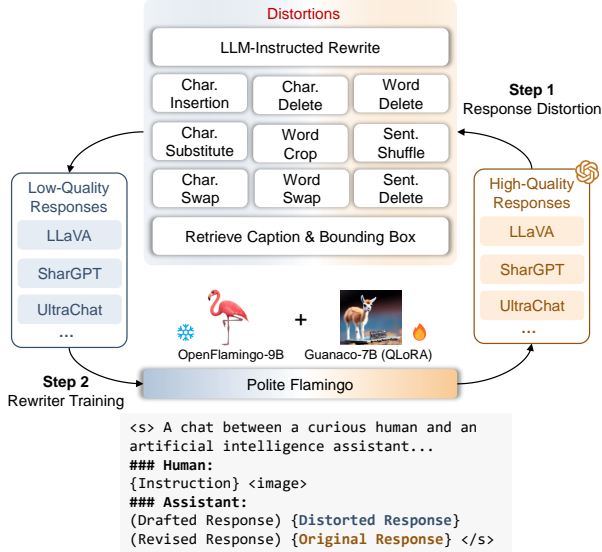


Figure 3. **Training pipeline of training Polite Flamingo.** We distort original high-quality responses into the corresponding low-quality version, then train a multi-modal LLM to predict the original response. This model is then used to rewrite raw annotations of a wide range of vision-language datasets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

or raw annotations) so that the train-test domain divergence is small and these denoising AutoEncoders can generalize well.

3.1. 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 are shown in Figure 4.

- **LLM-instructed Distortion.** Representative patterns of raw annotations include short answers (*e.g.*, VQA-v2 [32]), lacking punctuation or capitalization (*e.g.*, MS-COCO Captions [29]), not being coherent (*e.g.*, A-OKVQA [37]), etc., and we prompt an LLM (Guanaco [38]³) 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%. The distortion choices, which aim to further mimic the style of raw

³We used the QLoRA-based Guanaco language model [38], known for its superior performance (33B version, which has an average win rate of 97.8% against ChatGPT evaluated by GPT-4).

annotations, include capitalization modifications, inserting repetitions, using incorrect tenses, removing formatting, adding irrelevant information, etc. See Table 5 in the appendix for the detailed prompt structure.

- **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 NLPAUG⁴ 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 [9], 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.*, Flickr-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.

3.2. Source Datasets

When selecting the source datasets for training Polite Flamingo, we take into consideration the following three criteria. **1) Politeness:** The source datasets chosen should contain responses with a desired level of politeness. These responses will be directly learned by Polite Flamingo and subsequently transferred to the final model. **2) Multi-modality:** It is important for Polite Flamingo to leverage complementary visual information during the process of response rewriting. We expect it can provide necessary explanations for those short answers to ensure comprehensive and informative responses. **3) Diversity:** The training set must be sufficiently large to prevent the LLM-based Polite Flamingo from overfitting to specific patterns. According to the above criteria, we select three datasets to construct the training data for Polite Flamingo:

1. **LLaVA instructions [9]:** a multi-modal self-instruct dataset based on GPT-4, which is currently the only

⁴<https://github.com/makcedward/nlpaug>



LLM-instructed Rewrite

- 113k from LLaVA
- 76k from UltraChat and SharGPT

Instruction: Do the people in the image engaged in a conversation?

Original Response: Yes, the woman standing between the two men seems to be holding a conversation with them as they all share a laugh together.

↓ Distortion ↓

↑ Polite Flamingo ↑

Distorted Response: yes



Random Text Augmentations

- 77k from LLaVA

Instruction: What unique features can be found in this empty room?

Original Response: The empty room features various pieces of colorful furniture, including a tan couch and a large wood and glass armoire or trophy case...

↓ Distortion ↓

↑ Polite Flamingo ↑

Distorted Response: The empty room features various mp @ iNeTcVeLs of colorful furniture, including a tan couch and a large wood and glass armoire or trophy modern style, highlighted by a 5gPr% eeean accent ...



Retrieve Caption & Bounding Box

- 14k from LLaVA-Detail

Instruction: Analyze the image in a comprehensive and detailed manner.

Original Response: The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.

In the parking area, various luggage items are scattered around the vehicle...

↓ Distortion ↓

↑ Polite Flamingo ↑

Distorted Response: A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage ...

The followings are specific object locations within the image, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y.

person: [0.681, 0.242, 0.774, 0.694],

person: [0.63, 0.222, 0.686, 0.516]...

Figure 4. **Examples of our three response distortion strategies.** We transfer the style of LLM-generated high-quality instruction responses into an “impolite” version, approximating the distribution of raw vision-language dataset annotations. The converted data is used to train Polite Flamingo.

available LLM-generated multi-modal visual instruction tuning dataset. In this study, we assume that ChatGPT/GPT-4 produces responses that are considered satisfactory in terms of style⁵. Therefore, this dataset satisfies the criteria of both politeness and multi-modality.

2. **UltraChat [39]:** a large-scale text-only instruction dataset consisting of dialogues between two ChatGPT turbo APIs. Since the LLaVA instructions dataset contains only 117k data points, we select this dataset to compensate for the limited data diversity. UltraChat is generated by ChatGPT and has undergone post-processing and careful filtering [39], so we assume it provides satisfactory politeness.
3. **ShareGPT:** a dataset of conversations with ChatGPT that is shared by users and was used to train the Vicuna model. This dataset contains model responses to real-world user queries, resulting in good diversity. ShareGPT is also considered to be of high quality, as the resulting models (Vicuna) have shown superior performance [40].

3.3. Training a Rewriter

We gathered a total of 255k samples to train the Polite Flamingo. We initialize the model from OpenFlamingo-

⁵Since our methodology is data-driven, it is not limited to this particular style. Polite Flamingo can easily incorporate and adapt to other styles if we have access to sufficient high-quality data from other sources.

9B [20], and insert a LoRA [19] adapter (initialized from the QLoRA of Guanaco-7B [38]) into its LLaMA-7B [41] language model. We tune the LoRA weights only, and keep other parameters (*i.e.*, language model, ViT, perceiver, X-ATTN layers [7]) 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.

4. Scale Up Visual Instruction Tuning with Polite Flamingo

4.1. 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 [11, 14, 15], 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, including MS-COCO [29], Flickr-30k [42], TextCaps [43], Image2Paragraph [44], CC-3M [31], ELEVATER-IC [45], Spot-the-Diff [46], Image-

editing-requests [47], RefCOCOg [48], A-OKVQA [37], VQA-E [49], ScienceQA [50], VQA-v2 [32], GQA [51], OCR-VQA [52], PointQA [53], etc. We summarized detailed information in Section C and Table 6 in the appendix.

4.2. Filtering Strategies

Our rewriter, Polite Flamingo, is based on LLaMA-7B [41], 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. See Appendix D for implementation details.

4.3. U-shaped Multi-stage Visual Instruction Tuning

We first leverage the Polite Flamingo to rewrite the response of source datasets (Section 4.1), obtaining 1.17M samples. 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 [39], ShareGPT, OASST-1 [54], Alpaca-GPT-4 [55], GPTeacher, and InstructionWild [56]. 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 $10\times$ 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.

4.4. 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.

5. Evaluations

5.1. How Does Polite Flamingo Rewrite the Response?

5.1.1 Qualitative Evaluation

First, we present a qualitative analysis of Polite Flamingo’s rewriting. In Figure 5, we show representative examples of both good (upper) and bad (bottom) cases, and note how Polite Flamingo rewrites examples as expected and how it makes mistakes. Overall, Polite Flamingo successfully converts raw annotations into polite, rich, and coherent responses. From various examples, it is observed that it is capable of 1) integrating information from multiple captions and/or bounding boxes, 2) improving response coherency, and 3) generating complete sentences/paragraphs from short annotations, etc.

Good Cases. One interesting example is shown in the center of the upper half – the “Eurofighter Typhoon” from ELEVATER’s FGVC-Aircraft dataset. The source dataset provides external knowledge retrieved from Wikipedia, WordNet, and GPT-3, as knowledge augmentations. However, in this example, the original external knowledge is mismatched with the image due to word ambiguity (a type of aircraft vs. a climate concept). As Polite Flamingo is a multi-modal LLM that can observe both image and text, it recognized this mismatch and modified it to the correct version. Another example is shown on the right side of the Typhoon example (from OCR-VQA [52] dataset), in which

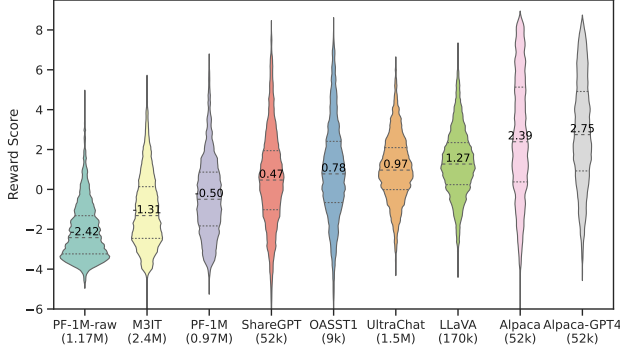


Figure 6. **Distribution of absolute reward model score of various instruction tuning datasets.** The median and quartile are also marked by dotted lines. Polite Flamingo boost the “politeness” of raw dataset annotations (leftmost) significantly.

the Polite Flamingo added the book title information to its rewritten answer. These examples illustrate the advantage of Polite Flamingo-based response rewriting in comparison with those ChatGPT-based ones (*e.g.*, in MIMIC-IT [15], M3IT [14], FuseCap [34], etc.). The multi-modality understanding ability of Polite Flamingo enables it to have a more comprehensive understanding of the instruction-response sample than the text-only rewriters.

Bad Cases. However, compared to ChatGPT-based rewriters, a major drawback of Polite Flamingo is its reliability – Polite Flamingo still makes some silly mistakes. In the bottom half of Figure 5, we show some representative examples of low-quality rewriting. Despite simple mistakes such as forgetting to generate `<EOS>` token thus producing endless repetitions, notable issues include changing the ground truth answer or adding hallucinated contents. It seems that sometimes Polite Flamingo prefers to believe its own visual perception rather than the provided ground truth, and its visual perception is not always accurate – possibly because the base model of Polite Flamingo, the OpenFlamingo-9B, is only trained on 15M image-text data thus produce less comprehensive visual representation alignment. These examples also demonstrate the necessity of post-processing and filtering.

5.1.2 Quantitative Evaluation

In addition to the above examples, we analyze the improvement of “politeness” 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 politeness⁶. In Figure 6, we plot the distribution of the scores of the reward model on a wide range of popular in-

struction tuning datasets⁷. It shows 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 M³IT [14] 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 [9] and Alpaca-GPT-4 [55]). 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.

5.2. Comparing Clever Flamingo with Existing Multi-modal LLMs

Now we turn to verify the performance of the Clever Flamingo trained with PF-1M, and compare it with other 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-preferred responses (*i.e.*, being polite). We first compare it with other models on image captioning and VQA tasks (Section 5.2.1), then we present the evaluation of multi-image reasoning tasks (Section 5.2.2), and finally, we analyze the politeness of these multi-modal LLMs (Section 5.2.3).

5.2.1 Image Captioning and VQA

Table 1 summarized the evaluation results comparing Clever Flamingo with other multi-modal LLMs on detailed image captioning (MS-COCO [29], TextCaps [43], and Image2Paragraph [44]) and visual question answering (OK-VQA [57], Visual-Spatial Reasoning [21], Grid-3D [58]). We use Rouge-L as the metric for captioning datasets and use an NLI model-based automated evaluator for VQA datasets (see Section F for more details). As our work is concurrent with InstructBLIP [11] and Otter [12], 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,

⁶Reward model: <https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2>

⁷10k samples are randomly drawn from each dataset. At the time of writing, another concurrent visual instruction tuning dataset MIMIC-IT [15] (which is used to train Otter) is not fully available.

Table 1. **Performance comparison of different multi-modal LLMs.** We use Rouge-L as the metric for detailed image description tasks (MS-COCO, TextCaps, and Image-to-paragraph), and we use an NLI-based evaluator for VQA datasets (OK-VQA, Visual-Spatial Reasoning, and Grid-3D). **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 ($\pm\Delta$) compares Clever Flamingo with Otter, which uses the same OpenFlamingo-9B as the base model.

Method	#Instructions	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
Otter	2.8M	ViT-L (0.3B)	OpenFlamingo-9B (15M)	13B	23.5	19.7	22.1	52.2	48.9	27.5
Clever Flamingo	1.0M	ViT-L (0.3B)		7B	22.6	19.7	22.4	28.7	28.7	13.5
$\pm\Delta$	-1.8M	-		7B	24.3	24.1	24.7	43.3	43.6	29.0
					+1.7	+4.4	+2.3	+14.6	+14.9	+15.5

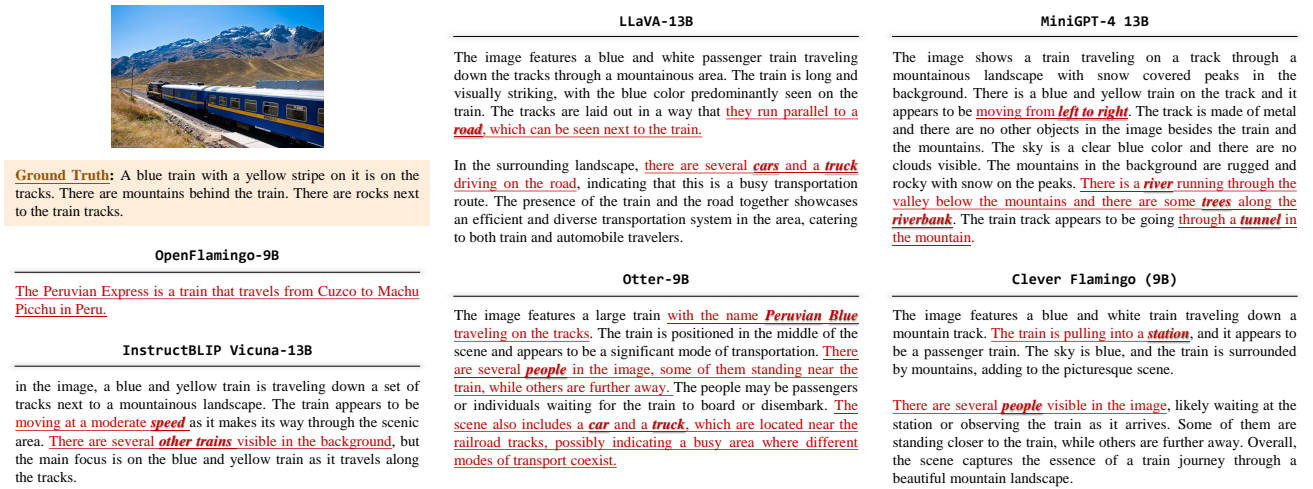


Figure 7. **All existing multi-modal LLMs exhibit severe hallucination problems.** We show a random testing example from the Image2Paragraph [44]. The hallucinated contents are marked with red.

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 \times smaller visual encoder, a lighter perceiver as the connector, and much less pre-training image-text data (15M) and training steps (single epoch)⁸ [20]. 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.

Hallucination Problem. Although Clever Flamingo yields notable improvement in image captioning tasks, it still exhibits severe object hallucination problems [59–61], the same as other existing multi-modal LLMs. In Figure 7,

we prompt existing multi-modal LLMs with the instruction “Give an elaborate explanation of the image you see”, using a random testing sample⁹ from the Image2Paragraph [44] dataset. As marked with red, all of the compared models hallucinated non-exist objects, such as road, cars, trucks, people, river, trees, tunnel, station, etc. This is a significant limitation faced by existing multi-modal LLMs, preventing them to be actually deployed in the real world. We also find that it is difficult to quantitatively verify the correctness of generation beyond object appearance (*e.g.*, “this is a scenic area”, “the train is visually striking”, “beautiful mountain landscape”, “this is a busy transportation route”, etc.), as we lack a dataset with rich fine-grained annotations of all information that can be inferred from the image.

⁸<https://laion.ai/blog/open-flamingo/>

⁹Not cherry/lemon-picked – it is the first image in our sampled validation set.

Table 2. **Multi-image reasoning tasks**. “STS” means semantic textual similarity. The lower bound performance 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).

Model	Spot-the-Diff		Image-editing		NLVR2	
	STS	Rouge-L	STS	Rouge-L	STS	Rouge-L
Lower Bound	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
Clever Flamingo	46.1	0.185	37.0	0.156	28.2	0.155
$\pm\Delta$	+6.6	+0.057	+3.9	+0.020	+16.7	+0.085

5.2.2 Multi-image Reasoning

Now we analyze the performance on multi-image reasoning tasks. We compare Clever Flamingo with Otter [12], 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 [46], a change captioning dataset for surveillance camera imagery, 2) Image-editing-requests [47], which requires the model to infer image editing requests (*e.g.*, Photoshop editing) given image pairs, and 3) Natural Language Visual Reasoning-2 (NLVR2) [62], 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 features¹⁰, to compare high-level semantics of answer and ground truth [63]. 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.

5.2.3 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 datasets¹¹. 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 8, we visualize the average win rate – the statistics 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,

¹⁰STS model: <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

¹¹See Section E in the appendix for details.

Model A	Raw Annotation	50.0	64.6	49.2	53.5	53.0	32.6	20.4	16.2	13.8
	MiniGPT-4-7b	35.4	50.0	36.6	37.5	37.3	21.6	16.0	10.5	8.9
	MiniGPT-4-13b	50.8	63.4	50.0	50.2	49.9	33.8	25.8	19.1	16.8
	InstructBLIP (Vicuna-7b)	46.5	62.5	49.8	50.0	49.0	35.7	26.7	21.4	19.7
	InstructBLIP (Vicuna-13b)	47.0	62.7	50.1	51.0	50.0	36.2	26.6	21.7	20.0
	Otter-9B	67.4	78.4	66.2	64.3	63.8	50.0	37.9	29.6	26.4
	Clever Flamingo	79.6	84.0	74.2	73.3	73.4	62.1	50.0	42.9	39.4
	LLaVA-7B	83.8	89.5	80.9	78.6	78.3	70.4	57.1	50.0	45.3
	LLaVA-13B	86.2	91.1	83.2	80.3	80.0	73.6	60.6	54.7	50.0
		Raw Annotation	MiniGPT-4-7b	MiniGPT-4-13b	InstructBLIP (Vicuna-7b)	InstructBLIP (Vicuna-13b)	Otter-9B	Clever Flamingo	LLaVA-7B	LLaVA-13B
		Model B								

Figure 8. **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.

except the LLaVA series. This is a direct result of the differences in instruction data, as in previous Figure 6, GPT-4 generated LLaVA dataset outperforms the PF-1M dataset in terms of reward score.

5.3. 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 9. 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 slight overfitting to the PF-1M data distribution, well stage 3 alleviates this problem.

The right side of Figure 9 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 per-

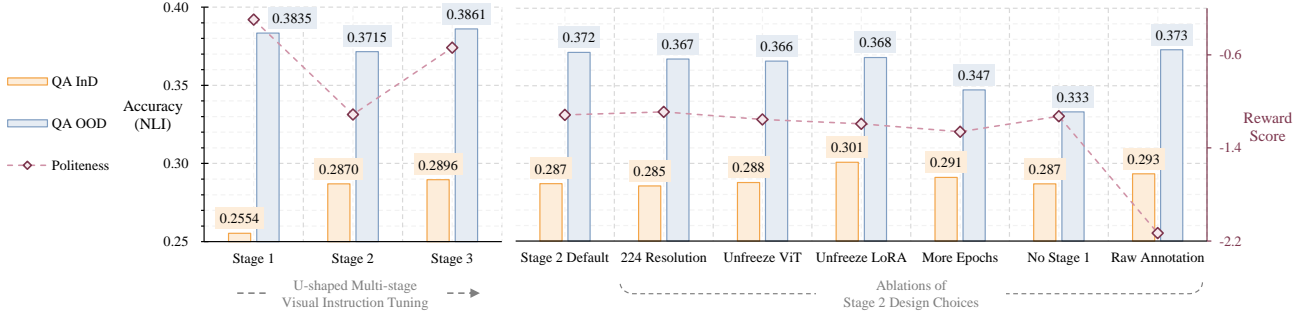


Figure 9. **Results of ablation experiments on U-shaped multi-stage visual instruction tuning (left) and design choices in stage 2 (right).** We calculate averaged NLI-based accuracy for **held-in QA datasets (QA InD)** and **held-out datasets (QA OOD)**. Note that all testing examples are unseen during training, and the difference between QA InD and QA OOD is the data domain distribution. We also report the **average reward score** to reflect the politeness of each alternative.

Table 3. **Comparison of response rewriting of two generations of Polite Flamingo.** The second generation is a politer and more active rewriter on A-OKVQA [37], but it fails to improve the CLIPScore on Conceptual Captions-3M [31].

Rewriter	A-OKVQA		CC-3M
	Avg. Reward	Unchanged	CLIPScore
Polite Flamingo Generation 1	-1.19	11.53%	23.88
Polite Flamingo Generation 2	-0.68	0.00%	23.25

formance, confirmed the hypothesize in [64]. **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.

5.4. The Second Generation of Polite Flamingo

As shown in Table 1 and Figure 9, we confirmed that Clever Flamingo has an improved visual perception and understanding ability through visual instruction tuning on PF-1M. We hypothesize these advantages might be transferred to benefit response rewriting, by tuning Clever Flamingo

to learn response rewrite. If the second generation of Polite Flamingo becomes a better rewriter, we may expect the subsequent second generation of Clever Flamingo could be further improved, and then a weakly supervised training loop become possible to be realized. To verify the possibility, we made an initial attempt by training and evaluating a second generation of Polite Flamingo. We use exactly the same training recipe as the first generation, except that we initialize the model from Clever Flamingo instead of OpenFlamingo-9B. After training, we use this second generation of Polite Flamingo to rewrite responses in A-OKVQA [37] and 20k samples from the CC-3M [31].

The results are shown in Table 3. We found that the second generation has a notable improvement (+0.51) in terms of average reward score. Additionally, the first generation of Polite Flamingo left 11.53% of samples as original and failed to make any revisions, while no sample remains unchanged by the second generation. The above observations demonstrate that the second generation of Polite Flamingo becomes a politer and more active rewriter.

However, the second generation failed to improve the CLIPScore of generated captions on CC-3M as expected. This is surprising as it seems to contradict our experimental results, where Clever Flamingo demonstrated a clear improvement over baselines. The most possible explanation for this phenomenon could be the rewriting style is limited by the training data distribution of Polite Flamingo (Section 3.2). Although it covers samples from multiple datasets, examples of describing images only appear in the LLaVA dataset, and there are only 23k samples for this type. It appears that our model overfits these 23k samples, as they are the *only* source to learn image captioning style throughout the whole process¹². This confirms our emphasis on the

¹²The captioning samples in PF-1M could not provide additional diversity that helps prevent overfitting, as they are also generated by the Polite Flamingo that learns to caption from the 23k samples only

importance of diversity when selecting training data for Polite Flamingo (Section 3.2), and reveals the urgent need for visual instruction tuning data of the detailed captioning type that is both high-quality and large-scale.

6. 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. Our methodology brings inspiration from denoising AutoEncoders, and the “noise” here is implemented by various text distortions that attempt to approximate the style of raw annotations to ensure generalization. Empirically, we implemented and trained the rewriter, and used it to build a large-scale visual instruction tuning dataset. Incorporating newly proposed U-shaped multi-stage visual instruction tuning and multi-turn augmentation, we derived a strong multi-modal LLM based on the dataset. We evaluate the resulting model on various tasks, and demonstrated its advantages in terms of both multi-modal understanding and response politeness.

References

- [1] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, “OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework,” in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 23318–23340, PMLR, 2022. [1](#)
- [2] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, “Unified-io: A unified model for vision, language, and multi-modal tasks,” *CoRR*, vol. abs/2206.08916, 2022. [1](#)
- [3] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, and J. Dai, “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*, pp. 16783–16794, IEEE, 2022. [1](#)
- [4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, “Training language models to follow instructions with human feedback,” in *NeurIPS*, 2022. [1](#), [20](#)
- [5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen, “A survey of large language models,” *CoRR*, vol. abs/2303.18223, 2023. [1](#)
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “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* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, PMLR, 2021. [1](#)
- [7] J. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, “Flamingo: a visual language model for few-shot learning,” in *NeurIPS*, 2022. [1](#), [5](#), [15](#)
- [8] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, “Language is not all you need: Aligning perception with language models,” *CoRR*, vol. abs/2302.14045, 2023. [1](#)
- [9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *CoRR*, vol. abs/2304.08485, 2023. [1](#), [3](#), [4](#), [8](#)
- [10] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigt-4: Enhancing vision-language understanding with advanced large language models,” *CoRR*, vol. abs/2304.10592, 2023. [1](#), [2](#), [3](#)
- [11] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” *CoRR*, vol. abs/2305.06500, 2023. [1](#), [2](#), [3](#), [5](#), [8](#)
- [12] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, “Otter: A multi-modal model with in-context instruction tuning,” *CoRR*, vol. abs/2305.03726, 2023. [1](#), [2](#), [3](#), [8](#), [10](#)
- [13] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lucic, M. Tschannen, A. Nagrani, H. Hu, M. Joshi, B. Pang, C. Montgomery, P. Pietrzyk, M. Ritter, A. J. Piergiovanni, M. Minderer, F. Pavetic, A. Waters, G. Li, I. Alabdulmohsin, L. Beyer, J. Amelot, K. Lee, A. P. Steiner, Y. Li, D. Keysers, A. Arnab, Y. Xu, K. Rong, A. Kolesnikov, M. Seyedhosseini, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, “Pali-x: On scaling up a multilingual vision and language model,” *CoRR*, vol. abs/2305.18565, 2023. [1](#)
- [14] L. Li, Y. Yin, S. Li, L. Chen, P. Wang, S. Ren, M. Li, Y. Yang, J. Xu, X. Sun, L. Kong, and Q. Liu, “M³it: A large-scale dataset towards multi-modal multilingual instruction tuning,” *CoRR*, vol. abs/2306.04387, 2023. [1](#), [3](#), [5](#), [8](#), [19](#), [20](#)
- [15] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, “MIMIC-IT: multi-modal in-context instruction tuning,” *CoRR*, vol. abs/2306.05425, 2023. [2](#), [3](#), [5](#), [8](#)
- [16] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, “Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models,” *CoRR*, vol. abs/2306.09265, 2023. [2](#)
- [17] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen, “Multimodal-gpt: A vision and language model for dialogue with humans,” *CoRR*, vol. abs/2305.04790, 2023. [2](#), [3](#)
- [18] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, “Towards a unified view of parameter-efficient transfer learning,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. [2](#)
- [19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022. [2](#), [5](#), [15](#)

- [20] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, “Openflamingo,” *Mar. 2023*, **2**, **5**, **9**, **15**
- [21] F. Liu, G. Emerson, and N. Collier, “Visual spatial reasoning,” *CoRR*, vol. abs/2205.00363, 2022. **2**, **3**, **8**, **19**
- [22] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, “Self-instruct: Aligning language model with self generated instructions,” *CoRR*, vol. abs/2212.10560, 2022. **3**
- [23] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, “A survey on multimodal large language models,” *arXiv preprint arXiv:2306.13549*, 2023. **2**
- [24] A. Zeng, A. Wong, S. Welker, K. Choromanski, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, “Socratic models: Composing zero-shot multimodal reasoning with language,” *CoRR*, vol. abs/2204.00598, 2022. **2**
- [25] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface,” *CoRR*, vol. abs/2303.17580, 2023. **2**
- [26] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, “MM-REACT: prompting chatgpt for multimodal reasoning and action,” *CoRR*, vol. abs/2303.11381, 2023. **2**
- [27] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao, “Llama-adapter: Efficient fine-tuning of language models with zero-init attention,” *CoRR*, vol. abs/2303.16199, 2023. **3**
- [28] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen, J. Tian, Q. Qi, J. Zhang, and F. Huang, “mplug-owl: Modularization empowers large language models with multimodality,” *CoRR*, vol. abs/2304.14178, 2023. **3**
- [29] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft COCO captions: Data collection and evaluation server,” *CoRR*, vol. abs/1504.00325, 2015. **3**, **4**, **5**, **8**, **17**
- [30] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *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* (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, eds.), pp. 1143–1151, 2011. **3**
- [31] P. Sharma, N. Ding, S. Goodman, and R. Soicrut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers* (I. Gurevych and Y. Miyao, eds.), pp. 2556–2565, Association for Computational Linguistics, 2018. **3**, **5**, **11**, **17**
- [32] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “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*, pp. 6325–6334, IEEE Computer Society, 2017. **3**, **4**, **6**, **18**
- [33] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, J. Shao, and W. Ouyang, “LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark,” *CoRR*, vol. abs/2306.06687, 2023. **3**
- [34] N. Rotstein, D. Bensaïd, S. Brody, R. Ganz, and R. Kimmel, “Fusecap: Leveraging large language models to fuse visual data into enriched image captions,” *CoRR*, vol. abs/2305.17718, 2023. **3**, **8**
- [35] L. Fan, D. Krishnan, P. Isola, D. Katabi, and Y. Tian, “Improving CLIP training with language rewrites,” *CoRR*, vol. abs/2305.20088, 2023. **3**
- [36] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, and J. Zhou, “Remotecap: A vision language foundation model for remote sensing,” *CoRR*, vol. abs/2306.11029, 2023. **3**
- [37] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Motlaghi, “A-OKVQA: A benchmark for visual question answering using world knowledge,” in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII* (S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13668 of *Lecture Notes in Computer Science*, pp. 146–162, Springer, 2022. **4**, **6**, **11**, **18**
- [38] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient finetuning of quantized llms,” *CoRR*, vol. abs/2305.14314, 2023. **4**, **5**, **15**
- [39] N. Ding, Y. Chen, B. Xu, Y. Qin, Z. Zheng, S. Hu, Z. Liu, M. Sun, and B. Zhou, “Enhancing chat language models by scaling high-quality instructional conversations,” *CoRR*, vol. abs/2305.14233, 2023. **5**, **6**
- [40] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *CoRR*, vol. abs/2306.05685, 2023. **5**
- [41] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” *CoRR*, vol. abs/2302.13971, 2023. **5**, **6**
- [42] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014. **5**, **17**
- [43] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, “Textcaps: A dataset for image captioning with reading comprehension,” in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12347 of *Lecture Notes in Computer Science*, pp. 742–758, Springer, 2020. **5**, **8**, **17**
- [44] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, “A hierarchical approach for generating descriptive image paragraphs,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3337–3345, IEEE Computer Society, 2017. **5**, **8**, **9**, **17**
- [45] C. Li, H. Liu, L. H. Li, P. Zhang, J. Aneja, J. Yang, P. Jin, H. Hu, Z. Liu, Y. J. Lee, and J. Gao, “ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models,” in *NeurIPS*, 2022. **5**, **18**
- [46] H. Jhamtani and T. Berg-Kirkpatrick, “Learning to describe differences between pairs of similar images,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), pp. 4024–4034, Association for Computational Linguistics, 2018. **5**, **10**, **18**
- [47] H. Tan, F. Dernoncourt, Z. Lin, T. Bui, and M. Bansal, “Expressing visual relationships via language,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (A. Korhonen, D. R. Traum, and L. Màrquez, eds.), pp. 1873–1883, Association for Computational Linguistics, 2019. **6**, **10**, **18**
- [48] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *Computer Vision - ECCV 2016*

- 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, *Proceedings, Part II* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), vol. 9906 of *Lecture Notes in Computer Science*, pp. 69–85, Springer, 2016. 6, 18
- [49] Q. Li, Q. Tao, S. R. Joty, J. Cai, and J. Luo, “VQA-E: explaining, elaborating, and enhancing your answers for visual questions,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), vol. 11211 of *Lecture Notes in Computer Science*, pp. 570–586, Springer, 2018. 6, 18
- [50] P. Lu, S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to explain: Multimodal reasoning via thought chains for science question answering,” in *NeurIPS*, 2022. 6, 18
- [51] D. A. Hudson and C. D. Manning, “GQA: A new dataset for real-world visual reasoning and compositional question answering,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709, Computer Vision Foundation / IEEE, 2019. 6, 18
- [52] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, “OCR-VQA: visual question answering by reading text in images,” in *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pp. 947–952, IEEE, 2019. 6, 18
- [53] A. Mani, W. Hinthorn, N. Yoo, and O. Russakovsky, “Point and ask: Incorporating pointing into visual question answering,” *CoRR*, vol. abs/2011.13681, 2020. 6, 18
- [54] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick, “Openassistant conversations - democratizing large language model alignment,” *CoRR*, vol. abs/2304.07327, 2023. 6, 15
- [55] B. Peng, C. Li, P. He, M. Galley, and J. Gao, “Instruction tuning with GPT-4,” *CoRR*, vol. abs/2304.03277, 2023. 6, 8
- [56] F. Xue, K. Jain, M. H. Shah, Z. Zheng, and Y. You, “Instruction in the wild: A user-based instruction dataset,” <https://github.com/XueFuzhao/InstructionWild>, 2023. 6
- [57] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, “OK-VQA: A visual question answering benchmark requiring external knowledge,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204, Computer Vision Foundation / IEEE, 2019. 8
- [58] J. H. Lee, M. Kerzel, K. Ahrens, C. Weber, and S. Wermter, “What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022* (L. D. Raedt, ed.), pp. 1039–1045, ijcai.org, 2022. 8
- [59] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, pp. 248:1–248:38, 2023. 9
- [60] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J. Wen, “Evaluating object hallucination in large vision-language models,” *CoRR*, vol. abs/2305.10355, 2023. 9
- [61] W. Dai, Z. Liu, Z. Ji, D. Su, and P. Fung, “Plausible may not be faithful: Probing object hallucination in vision-language pre-training,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023* (A. Vlachos and I. Augenstein, eds.), pp. 2128–2140, Association for Computational Linguistics, 2023. 9
- [62] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, “A corpus for reasoning about natural language grounded in photographs,” in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (A. Korhonen, D. R. Traum, and L. Márquez, eds.), pp. 6418–6428, Association for Computational Linguistics, 2019. 10
- [63] N. Reimers and I. Gurevych, “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, 11 2019. 10
- [64] Y. Liu, Z. Li, H. Li, W. Yu, M. Huang, D. Peng, M. Liu, M. Chen, C. Li, L. Jin, and X. Bai, “On the hidden mystery of OCR in large multimodal models,” *CoRR*, vol. abs/2305.07895, 2023. 11
- [65] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, “Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 60, pp. 1–19, 2022. 17
- [66] X. Lu, B. Wang, X. Zheng, and X. Li, “Exploring models and data for remote sensing image caption generation,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 56, no. 4, pp. 2183–2195, 2018. 17
- [67] A. Shtedritski, C. Rupprecht, and A. Vedaldi, “What does CLIP know about a red circle? visual prompt engineering for vlms,” *CoRR*, vol. abs/2304.06712, 2023. 18

Appendix

A. Implementation Details

We implemented our approach on the OpenFlamingo codebase [20]¹³, which is an open-source re-implementation of DeepMind’s Flamingo [7]. Our training was performed on a single node machine with 8 NVIDIA A100 (40GB) GPUs. To accommodate memory limitations, we utilized BF-16 precision for training and inference of Polite/Clever Flamingo. Detailed settings and hyperparameters are summarized in Table 4.

Table 4. Training details of Polite Flamingo and Clever Flamingo.

	Polite Flamingo	Clever Flamingo Stage 1	Clever Flamingo Stage2	Clever Flamingo Stage 3
Tunable Modules	LoRA	LoRA	Perceiver, XATTN	LoRA
Tunable Parameters	0.29B	0.29B	0.1B	0.29B
Number of Samples	255k	772k	1.07M	772k
Epochs	3	1	3	1
Learning Rate	1e-4	1e-4	1e-4	1e-5
Batch Size	256	256	1024	256
Context Length	1024	1024	196	1024
Maximum Images	-	10	3	10
Training Time (hours)	11.8	11.8	9.5	11.5

Model Architecture. Polite/Clever Flamingo is initialized from the OpenFlamingo-9B (v1) checkpoint and inherits the architecture from the base model. It comprises a (vanilla, not instruction-tuned) LLaMA-7B language model, a ViT encoder from OpenAI’s CLIP (ViT-Large-14), a perceiver, and interleaved XATTN layers inserted into the language model.

- **Language Model:** We insert a LoRA [19] adapter into the language model (for both self-attention and FFN), initialized from QLoRA-Guanaco-7B [38]. The LoRA adapter is trained on the OASST-1 instruction dataset [54] and has a rank of 64.
- **ViT Encoder:** OpenFlamingo-9B uses the ViT-Large-14 as the vision encoder, taking image inputs with a resolution of 224×224 . We substitute it with ViT-Large-14@336pix, which undergoes an additional CLIP pretraining epoch with a resolution of 336×336 . Empirically, we observed that the representation distribution does not shift significantly compared to the 224×224 version, enabling seamless substitution.
- **Perceiver:** The perceiver resampler takes patch tokens from ViT as input and pools them to 64 tokens. Its size is roughly equivalent to one layer of ViT.
- **XATTN Layers:** Following Flamingo [7], XATTN layers are inserted into the LLaMA-7B every 4 LM layers. XATTN consists of cross-attention and FFN. When referring to “unfreezing XATTN,” we mean unfreezing the weights of cross-attention only while keeping the FFN frozen.

Multi-turn Augmentation. During the training of Clever Flamingo, when loading each instruction sample, we randomly draw samples from the dataset to fill the $\langle \text{PAD} \rangle$ tokens. These samples are appended to the first sample to simulate subsequent rounds of conversation. No system message (*i.e.*, “A chat between a curious human and an artificial intelligence assistant...”) is added for later turns. The end-of-sentence token $\langle \text{EOS} \rangle$ is appended to each response, and the loss is only calculated for the AI assistant response parts (between each “### Assistant: ” and the $\langle \text{EOS} \rangle$ of the corresponding response). To obtain a loss mask (for setting the label index to -1 in language modeling), per-turn tokenization is required. However, we empirically found that this does not affect training efficiency.

B. Polite Flamingo Training Data

A total of 255k samples were used for training Polite Flamingo, including:

1) **LLM-instructed Distortion:** The prompt structure for LLM-instructed rewrite (Section 3.1) is shown in Table 5. Using this prompt structure, Guanaco-33B generated 133k multi-modal (LLaVA) and 76k text-only (UltraChat + ShareGPT) distortion samples.

¹³https://github.com/mlfoundations/open_flamingo

Prompt for LLM-instructed Distortion

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user’s questions.

Human:

{Instruction}

Assistant:

{Original Response}

Human:

Your reply’s style, tone, and politeness are excellent, and the content is very detailed. However, now I would like you to summarize the previous response, keeping only the most crucial information and removing all other less important content. I want a concise, straightforward reply without any redundancy. If you find that the overall quality of your response dropped, don’t worry, it’s fine. Note that, please do not add anything after giving me your rewritten response.

Assistant:

Sure. I have rewritten my last response to a much shorter and more concise version, covering only the key information. I pretend to be a cold-hearted, non-talkative, socially inept robotic assistant to respond to your request. {Distortion} The following is the as-short-as-possible, low-quality, highly-compressed, rewritten version of my previous response, and I will not add more content after finishing this response: ”

Distortion Choices:

- Additionally, I have **removed all the punctuation marks and capitalization** in my response.
- To make my response more unnatural, I have added a little amount of **typos and spelling mistakes**.
- I have also added some **grammatical errors** to my response.
- Moreover, **random words and sentences** have been removed from my response.
- In addition, all letters in my response have been converted to **uppercase**.
- In addition, all letters in my response have been converted to **lowercase**.
- Furthermore, I have replaced certain words with their **synonyms** in my response.
- Additionally, I have inserted unnecessary **repetition** in my response.
- To make my response less **coherent**, I have rearranged the sentence structure.
- I have deliberately used **incorrect tenses** and **verb conjugations** in my response.
- Moreover, I have introduced unnecessary **verbosity** in my response.
- I make my response **as short as possible** by removing all unnecessary words and sentences.
- I have kept **only the essential information** and separated them by commas.
- I have removed any decorative **formatting or styling**, which may affect the **readability** of my response.
- I have rewritten the sentences and replaced words with their **synonyms**.
- I have reversed the **order of sentences**, presenting information from back to front.
- I made my response sounds more **unprofessional and casual**.
- Furthermore, I have made the language more **complex and sophisticated** in my response.
- To create **ambiguity**, I have added multiple interpretations in my sentences.
- Additionally, I have used unconventional **metaphors and analogies** in my response.
- To lower the quality of my response, I have added some **irrelevant information**.
- I picked one sentence from my response and **repeated** it multiple times, each time with a slight change.
- Now I use only five words to **summarize** my response.
- I made some modification to make my response less **coherent** and more **unnatural**.

Table 5. **Prompt for LLM-instructed distortion.** We prompt LLM to translate the style of the original high-quality response into the “impolite” version, approximating the distribution of raw annotations in vision-language datasets.

2) Random Text Augmentations: We utilized the NLPAUG library for character-level, word-level, and sentence-level text augmentation. For character-level augmentation, we randomly selected an operation from character insertion, substitution, swapping, and deletion. Word-level augmentation operations included swapping, cropping, and deletion. Sentence-level augmentation involved randomly dropping sentences or shuffling their order. A total of 77k samples were generated using this method.

3) Retrieve Captions & Bounding Boxes: We obtained 14k samples of this type, which are non-overlapping with the LLM-instructed Distorted LLaVA-detailed-23k samples.

C. Clever Flamingo Training Data

We have provided a summary of the detailed composition of PF-1M in Table 6. Please note that "Adopted Samples" does not indicate the full training set size for all datasets, as Polite Flamingo was not applied to rewrite the entire dataset. Additionally, during the filtering step, a proportion of samples were removed.

Table 6. Details of the PF-1M dataset.

Category	Dataset	Adopted Samples	Description
Image Captioning	MS-COCO-2014	59,670	Image in MS-COCO Caption dataset has 5 human-generated captions.
	Flickr-30k	31,695	Dataset containing 31,000 images from Flickr with 5 reference sentences.
	TextCaps	69,703	Dataset for image captioning with reading comprehension.
	Image2Paragraph	7,954	Dataset with images from Visual Genome, each containing one paragraph.
	Conceptual Captions 3M	67,025	Google's Conceptual Captions dataset with millions of images and descriptions.
	RefCOCOg	8,103	Large-scale dataset for referring expressions based on MS-COCO.
	RET-3	13,551	Collection of image-text datasets (RSICD, RSITMD, UCM) introduced in RemoteCLIP.
	DIOR	5,907	Large-scale benchmark for object detection in Optical Remote sensing images.
	DOTA	1,733	Large-scale dataset for object detection in aerial images.
	HRRSD	5,898	Large-scale high-resolution remote sensing object detection dataset.
Image Classification		21,772	ELEVATER-IC benchmark collection for language-image models on image classification.
		ELEVATER Subset	# Classes (32 images are randomly sampled for each class)
		FER 2013	7
		CIFAR-10	10
		EuroSAT	10
		MNIST	10
		VOC 2007	20
		Oxford-IIIT Pets	37
		GTSRB	43
		Resisc-45	45
		Describable Textures	47
		CIFAR-100	100
		FGVC Aircraft	100
		Food-101	101
		Caltech-101	102
		Oxford Flowers 102	102
		Stanford Cars	196
Change Captioning	Spot-the-Diff	6,787	Dataset consisting of surveillance image pairs with annotations stating the differences.
	Image-Editing-Requests	2,747	Dataset with real image pairs and corresponding editing (<i>e.g.</i> , Photoshop) instructions.
	LevirCC-Captions	6,761	Large-scale dataset with pairs of bimodal RS images and sentences describing differences.
VQA with Rational	A-OKVQA	14,868	Crowdsourced VQA dataset requiring commonsense and world knowledge.
	VQA-E	65,133	Dataset for Visual Question Answering with Explanation.
	ScienceQA	4,596	Benchmark dataset for multimodal multiple-choice questions with lectures and explanations.
VQA without Rational	VQA-v2	210,743	Visual Question Answering (VQA) v2.0 dataset with open-ended questions about images.
	CLEVR	26,390	Synthetic Visual Question Answering dataset with 3D-rendered objects.
	GQA	223,244	Large-scale visual question answering dataset with real images from Visual Genome.
	TextVQA	30,056	Dataset for visual reasoning based on text in images.
	OCR-VQA	24,164	Dataset with question-answer pairs about book cover images.
	PointQA	67,282	Datasets for Visual Question Answering with pointing to objects in images.
Total		975,782	A combination of the above datasets.

- **Image Captioning:** With "Retrieve Caption & Bounding Box" distortion, the Polite Flamingo learned to integrate information from given multiple captions, bounding boxes, and its own visual perceptions, into several paragraphs of detailed captions. Leveraging this capability, we feed the MS-COCO [29], Flickr-30k [42], TextCap [43], and several datasets for earth observations [65, 66] to Polite Flamingo. Additionally, we introduce the Image2Paragraph [44] dataset, which offers comprehensive information coverage but lacks language coherence. We also incorporate the ConceptualCaptions-3M [31] dataset sourced from the web, which introduces further diversity to the captioning data. Recent studies have

demonstrated that CLIP models are capable of recognizing visual prompts, such as a red circle marked in an image [67]. Inspired by this, we adopt the RefCOCOg [48] dataset, converting the region of interest into annotations (colored bounding boxes or circles) in the image. Then, we accordingly set the instruction to “Describe the object inside this green bounding box.” for generating region-specific captions.

- **Image Classification:** ELEVATER-IC [45] is a diverse collection of image classification datasets, covering more than 1k visual concepts distributed in various domains. We introduce this dataset to enhance the fine-grained visual recognition capabilities. We simply set the instruction to “What is this?”, and use the prompt template originally for CLIP-based zero-shot classification (*e.g.*, a photo of a {class name}) to format the response. Furthermore, ELEVATER-IC provides additional external knowledge associated with each class, sourced from Wikipedia, WordNet, and GPT-3. We include this complementary information in the response to enrich the provided answer.
- **Change Captioning:** Existing change captioning models often require specific design such as complex attention mechanism. The emergence of multi-modal LLM, which is trained on interleaved image-text corpora and is able to process multiple images, makes it possible to solve change captioning more elegantly. To explore this potential, we adopt several change captioning datasets, such as Spot-the-Diff [46], to verify this potential. Additionally, we introduce the image-editing-requests [47], a dataset of image editing (*e.g.*, PhotoShop) requests collected from forums, to test higher-level comparison capability beyond just object appearance.
- **VQA with Rational:** In several VQA datasets, such as A-OKVQA [37], VQA-E [49], and ScienceQA [50], annotations of “explanation” or “rationale” are provided in addition to the answer. These contents offer valuable information for training a visual assistant AI. However, the coherence and readability of these rationale annotations are suboptimal. We introduce these datasets to Polite Flamingo for rewriting, aiming to enhance the clarity and coherence of the provided rationales.
- **VQA without Rational:** This group of datasets, including VQA-v2 [32], GQA [51], and OCR-VQA [52], have a larger scale in general. However, the answer annotations in these datasets typically comprise only a few words. We incorporate these datasets into Polite Flamingo to enable the generation of complete sentences for the provided answers. In line with the region captioning dataset, we include the PointQA dataset [53], which comprises question-answer pairs related to a specific point of interest in an image. To facilitate understanding, we mark the corresponding point with colored arrows based on the corresponding point coordinates in the image.

D. Filtering Strategies

Figure 10 shows our filtering pipeline to guarantee the quality of Polite Flamingo rewritten response and remove potential hallucinations. First, we introduce the length filter that excludes too-short or too-long responses. Then, we apply a change filter that removes responses that have not been rewritten – the underlying assumption is that the style of raw dataset annotation is undesired. Although these filters can remove many apparent low-quality samples, they cannot understand the semantics of the response and cannot identify hallucinated contents. To address this issue, we introduce several model-based filters, including a Semantic Textual Similarity (STS) model-based filter, Natural Language Inference (NLI) model-based filter, and a CLIPScore-based hallucination filter.

- **Semantic Textual Similarity (STS)-based Filter for Captioning Datasets:** We used a Sentence Transformer to analyze the semantic similarity between the original captions and rewritten captions. The Sentence Transformer we used is based on MPNet, and is trained with over 1 billion annotated sentence pairs¹⁴. We calculate the cosine distance between the sentence representation of original captions and their rewritten version, and remove the sample that scores below a threshold of 0.40.
- **CLIPScore-based Paragraph Filter for Captioning Datasets:** As LLaVA-detailed-23k is the only source that provides style reference of detailed image description in the training data of Polite Flamingo, it perfectly fits the style of this data. In this dataset, GPT-4 prefers to divide the visual contents into two paragraphs, and those second paragraphs usually start with “In addition/In the background, there are some ...”. Unfortunately, when the Polite Flamingo attempts to generate such a second paragraph, hallucinations are often been introduced, possibly due to the imperfect representation of the base model. To solve this problem, we calculate per-paragraph CLIPScore¹⁵, then remove the paragraphs with a CLIPScore lower than a threshold of 17.0.

¹⁴<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

¹⁵CLIPScore model: <https://huggingface.co/openai/clip-vit-large-patch14-336>

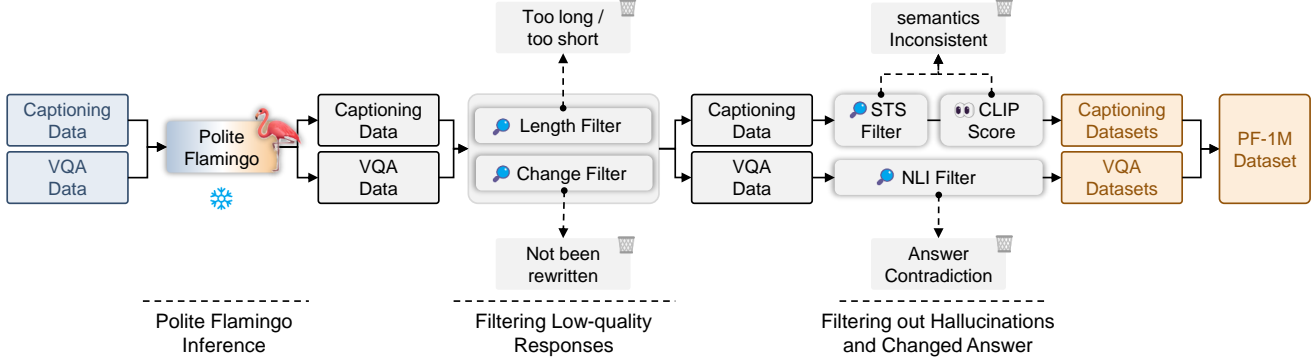


Figure 10. Filtering pipeline for Polite Flamingo written responses.

- **Natural Language Inference (NLI)-based Filter for VQA Datasets:** Occasionally, Polite Flamingo changes the original answer to another one during rewriting responses for VQA datasets – it trusts its own visual perception and its own thinking instead of the original answer. Possible reason includes imperfect representation, limited capacity of the 7B model, lacking certain regularization or sufficient data during its training process. To remove these samples, we employed an NLI model¹⁶, which is trained on SNLI and MultiNLI dataset and achieves 90.04% accuracy on MNLI mismatched set, to filter out rewritten answer that contradicts the original answer.

E. Evaluation Data

Table 1 and Table 2 benchmarks Clever Flamingo with other multi-modal LLMs on captioning and VQA datasets. For COCO (2014) dataset, we randomly drew 5k samples from its validation split. For TextCaps, Img2P, OK-VQA, Grid-3D, and NLVR2 datasets, we randomly drew 3k samples. Validation splits of VSR, Spot-the-Diff, and Imgae-editing-requests have fewer than 3k samples, so we use all available samples. The number of testing samples is limited due to the auto-regressive text generation of multi-modal LLMs being time-consuming.

Figure 8 presents the win rate comparison on 52k samples, which are sourced from various vision-language downstream datasets, including IconQA, VQAv2, OK-VQA, TextVQA, ScienceQA, VQA-E, ChartQA, GQA, OCR-VQA, A-OKVQA, AI2D, CLEVR, ELEVATER, VSR, and Grid3D. We adopt this wide collection to ensure the diversity of queries. Ablations in Figure 9 also adopt these datasets, and we further divide them into in-domain datasets and out-domain datasets, depending on whether it appears in PF-1M.

F. Automated Evaluators

NLI-based VQA Accuracy Evaluator. We utilized an NLI-based evaluator to benchmark multi-modal LLMs on VQA datasets. This evaluator is also based on the Sentence Transformer model `nli-deberta-v3-base`. The NLI model compares the model’s response and the ground truth answer with the prompt “{model answer}” is the answer to the question: “{question}” and “{ground truth}” is the answer to the question: “{question}”. An “entailment” output is considered a successful prediction. Compared to traditional evaluation methods such as exact match counting or the Rouge-L metric [14], our NLI-based evaluator is capable of capturing and comparing the semantic information of ground truths and model predictions more effectively. Additionally, compared to GPT-4-based evaluations [21], our NLI-based approach is more cost-effective, allowing us to scale up the validation sample size and obtain more robust results.

To validate the reliability of this model-based evaluator, we conducted a human evaluation. We randomly selected 600 samples from the evaluation data (Section E), which included 200 samples from OK-VQA, 100 samples from VSR, 100 samples from Grid-3D, and 200 samples from A-OKVQA, GQA, CLEVR, ChartQA, OCR-VQA, TextVQA, VQA-E, and VQAv2 (25 samples from each). Two human annotators were hired, with each annotator reviewing 300 out of the 600 samples. Afterward, cross-validation was performed, and any inconsistent annotations were modified based on a consensus reached through discussion.

¹⁶NLI model: <https://huggingface.co/cross-encoder/nli-deberta-v3-base>

For each of the 600 QA samples, we presented images, questions, ground truth answers, and model responses from 5 multi-modal LLMs. The annotators were asked to determine whether each model response falls into:

1. **Matched**: the model answer contains the ground truth and does not conflict with it.
2. **Correct**: the model answer does not match the ground truth, but it is still a valid and correct answer to the question.
3. **Failed**: the model answer neither matches the ground truth nor is a valid/correct answer.
4. **Uncertain**: it is not possible to determine whether the model answer is valid/correct.

We compared the human annotations with the results of the model-based evaluation as shown in Figure 11. The NLI-based evaluation accurately reflects the ranking of matched predictions. In contrast, the Rouge-L-based evaluator (as adopted in [14]) suggests that MiniGPT-4 is better than Otter and matches LLaVA, which significantly contradicts the human annotation results. Another observation is that the annotated ground truths in vision-language datasets are not the only valid ground truths, as there are clear gaps between “matched” predictions and “correct” predictions.

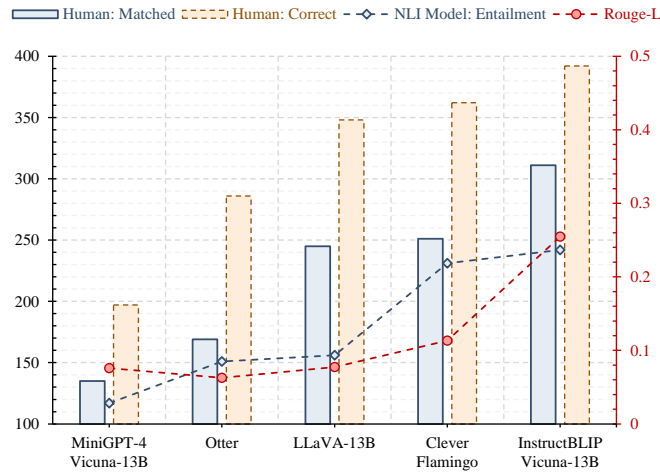


Figure 11. Meta-evaluation of NLI-based QA evaluator.

Reward Model-based Human Preference Evaluator. For the evaluation of politeness (i.e., human preference), we utilized a reward model¹⁷. This reward model was trained on various datasets, including **WebGPT Comparison**, **Summarize-from-Feedback**, **synthetic-instruct-gptj**, and **Anthropic-RLHF**. It achieved validation accuracies of 61.13%, 72.23%, 99.94%, and 55.62% on these datasets, respectively. This evaluation method is fair as none of the compared multi-modal LLMs involve any RLHF [4] process. We requested human annotators to rank the model responses of the 600 samples based on the following criteria:

1. Assuming all model responses are accurate and error-free, the preference ranking here does not consider the correctness of the answers.
2. Has the model accurately understood the question? Can the model’s response effectively answer the question?
3. Is the capitalization and punctuation in the model’s response accurate? Is the response coherent?
4. Is the length of the model’s response reasonable? Is it too short or excessively redundant/verbose?
5. As an AI assistant, does the tone of the model’s response come across as polite and align with user preferences?

We calculate the accuracy of the reward model in ranking pairs consistently with human annotations, excluding pairs labeled as “equally preferred.” The average accuracy is 70.0%, which is similar to the accuracy achieved in WebGPT Comparison, Summarize-from-Feedback, and Anthropic-RLHF. This demonstrates that the reward model effectively reflects user preferences.

¹⁷Reward model: <https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2>