ADAPTING UNI-MODAL LANGUAGE MODELS FOR DENSE MULTI-MODAL CO-REFERENCE RESOLUTION USING PARAMETER AUGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The context of modern smart voice assistants are often multi-modal, where images, audio and video content are consumed by users simultaneously. In such a setup, co-reference resolution is especially challenging, and runs across modalities and dialogue turns. We explore the problem of multi-modal co-reference resolution in multi-turn dialogues and quantify the performance of multi-modal LLMs on a specially curated dataset of long, image-interleaved conversations between a voice assistant and a human for a shopping use case. We propose and evaluate a custom architecture for multi-modal embedding alignment using a novel parameter augmentation technique. Our proposed Parameter Augmented LLM approach shows a 4.9% absolute F1 improvement above a baseline while reducing the number of parameters being trained by 13.3% for a complex co-referencinging task on a multi-turn shopping dataset.

1 INTRODUCTION

Large Language Models' (LLMs) success on text-only language tasks have been naturally followed by the question of whether they can generalize well to multi-modal tasks. To this end, several adaptation approaches have been proposed in the literature and tested on multi-modal tasks such as Image Captioning, Visual Question Answering (VQA), Image Selection, Image Instruction Following etc. Such tasks have a logical separation between image & text sequences, and the image-text input appears in a non-interleaved manner.

Compared to tasks such as VQA, there has been fewer work on multi-modal co-reference resolution in multi-turn dialogues that have interleaved images and text. Furthermore, most multi-modal dialogue datasets are based on one or two images, so most recent related work has been limited to this setup (Zang et al., 2021; Kottur et al., 2021). In this work, we go beyond this and leverage a multi-modal dataset with long dialogues ranging between 5 - 66 utterances that also has multiple images that are heavily co-referenced.

We propose a novel and efficient technique to convert a uni-modal LLM to a multi-modal one and test it on image selection and image retrieval tasks in a conversational setting. A sample dialogue is shown in Figure 1, where we color code the complex co-references that happen in the multi-turn dialog to give a sense of the difficulty of the task. Our approach tackles the co-reference problem across modalities by preserving the interleaved nature of multi-modal sequences and at the same time embedding text and images in a common space.

To quantify the performance of our proposal, we use an image selection task and an image retrieval task that occur within a multi-turn conversation as proxy for co-reference resolution. The image selection task involves a user making a visual attribute based referring utterance to select a specific image among many images on display (e.g., "select the red dress with orange stripes"). Here, we generate the position of the image that best suites a criterion using the Multi-Modal Context Carryover (MMCC) dataset (Wanigasekara et al., 2022; 2023). For image retrieval, we determine the images that are relevant at the final turn of a multi-modal multi-turn dialogue after curating the Multi-Modal Domain-Aware dataset (Saha et al., 2018) as depicted in Figure 1.

Figure 1: An example of a multi-turn dialogue with multi-modal co-referencing. The co-references are color coded and shown by arrows.

The summary of our contributions is as follows: (1) We propose and evaluate a parameter augmentation technique for transformer architectures and demonstrate a use case in adapting LLMs for multi-modal contexts. (2) We propose a novel approach for embedding alignment that preserves the spatial attributes of multi-modal interleaved data during processing. (3) We propose a cross modality normalization that is based on the Root Mean Square Layer Normalization (Zhang & Sennrich, 2019).

Our approach demonstrates an improvement over the best baseline (Accuracy=$+1.19\%$, F1=$+4.9\%$) as seen in Table 3 for the image retrieval task, implying better multi-modal co-reference resolution.

2 Related Work

There have been several elaborate image-text models over the years, such as CLIP (Radford et al., 2021) and BLIP models (Li et al., 2022; 2023). These separately train image and text encoders then train a text decoder to generate text conditioned on both the image and the text. This is achieved through some form of embedding alignment technique, e.g., CLIP (Radford et al., 2021) uses contrastive learning, BLIP (Li et al., 2022) uses cross attention while BLIP-2 (Li et al., 2023) uses a querying transformer.

Such models have been limited by the scale of available multi-modal data, which is far less than the currently available uni-modal data. To bridge this gap, large multi-modal datasets such as LAION (Schuhmann et al., 2022) have been proposed, but these are still significantly less than uni-modal datasets. For example LAION provides 2B English image-caption pairs but LLMs (Touvron et al., 2023a;b) are trained on trillions of text tokens. This has prompted the community to explore more on embedding alignment techniques, and so bring the best of Vision Models (VMs) and language models (LLMs) together to solve multi-modal problems. As demonstrated by MACAW (Lyu et al., 2023) and AnyMal (Moon et al., 2023), these same techniques generalize beyond image and text to audio, IMU sensor data and other modalities of data (Driess et al., 2023).

2.1 Multi-Modal Alignment Approaches

Image-text Alignment can be classified as either natural language alignment or embedding alignment. Natural language alignment between vision and language foundation models consists of first representing the vision input as text using an image-text model (such as CLIP, BLIP, and BLIP-2)

then processing the unified text using a language model (Guo et al., 2023; Wu et al., 2023). This has shown to have zero-shot capabilities, but can be limited because of its discrete nature. To overcome this, Visual ChatGPT (Wu et al., 2023) combines 22 vision foundation models for different vision tasks and a prompt manager that determines how the vision foundation models are used, a complex and resource-intensive setup.

Embedding alignment employs neural approaches to translate the embeddings of the vision foundation model to the embedding space of the language model. This approach can be robust, but does not have zero-shot capabilities unless pretrained on a multi-modal dataset first. To achieve such an alignment, Flamingo models (Alayrac et al., 2022; Awadalla et al., 2023) use a cross attention and contrastive learning objective, achieving state-of-the-art performance in several multi-modal tasks. Mini-GPT4 (Zhu et al., 2023) uses the querying transformer previously introduced in BLIP-2 (Li et al., 2023). Alternatively, one can use convolution and linear layer (Koh et al., 2023a; Lyu et al., 2023) with or without a separate modality encoder (Lyu et al., 2023; Moon et al., 2023; Koh et al., 2023b).

There are currently closed-source pipelines (OpenAI, 2023; Yang et al., 2023) that perform a similar multi-modal co-reference resolution task as ours. Given that they are closed-source and have the possibility of being a multi-modal mixture of experts setup, we do not directly compare them with our work.

2.2 Multi-Modal Co-Reference Resolution

To resolve the co-references in scene understanding multi-modal context, (Lee et al., 2022) proposes a binary classifier head on top of a transformer that predicts whether a current utterance mentions an object. To avoid a separate module for co-referencing, SHIKRA (Chen et al., 2023) instead uses numerical and language representation to point to specific parts of a scene, and this is infused in the text. A more systematic approach is to use a graph to relate visual and text entities (Guo et al., 2022). We instead choose to leverage LLM attention layers to learn multi-modal co-reference resolution because it will generalize over more contexts.

3 Our Approach

3.1 Motivation

In the techniques discussed previously (Alayrac et al., 2022; Awadalla et al., 2023; Zhu et al., 2023; Lyu et al., 2023; Koh et al., 2023a) there is a logical separation of input from the different modalities, even though the model may accept interleaved multi-modal input. For instance, cross attention uses one modality as attention query and another modality as attention key whereas the querying transformer learns queries from one modality then feeds it through self and cross attention layers to the other modality. We argue that such a logical separation, though sufficient for types of tasks where modalities are separate e.g. VQA, Image Captioning etc., is suboptimal for a multi-modal co-reference resolution. This is in line with findings from (Koh et al., 2023b) who observed poor performance when performing image retrieval task over multiple co-referenced images. We test out this hypothesis with OpenFlamingo (Awadalla et al., 2023) and our proposed Parameter Augmented LLM approach, which preserves the sequence of the multi-modal information during processing.

3.2 Problem Formulation

In our setup, a multi-modal dialogue $D := \{(U_i, S_i, I_i)\}_{i=1}^{s}$ contains s turns, each of them composed of a user textual utterance U_i, the system answer S_i, and the images I_i.

Due to the nature of the chosen datasets (i.e. shopping context), at each turn, the images I_i are interleaved within the system utterance, while the user utterance is fully uni-modal. Note however that both the user and the system can reference textual or image entities from past turns, requiring multi-modal co-reference resolution. An example of such an interaction is shown in Figure 1.

In what follows, we refer to token and image embeddings as $\mathbf{x}^t$ and $\mathbf{x}^v$. Our approach relies on augmenting a pre-trained LLM $h_\theta(\mathbf{x}^t)$ with frozen parameters θ and hidden dimension d_{llm}. We denote their augmented counterparts with a hat superscript: the augmented LLM is noted $h_{\theta,\hat{\theta}}(\mathbf{x}^t)$, with

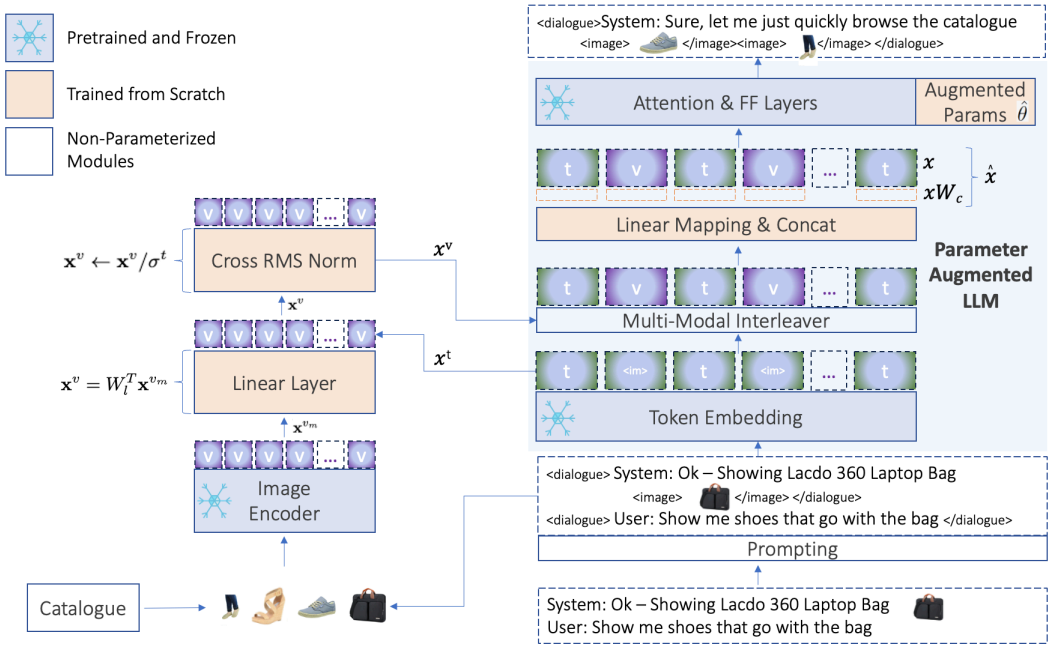

Figure 2: LLM-Agnostic Architecture for Parameter Augmentation. Boxes with t and v refers to text and image embeddings, respectively. We optimize the linear layer, cross RMS normalization module, and the augmented parameters, the rest of the LLM remains frozen. The Multi-Modal Interleaver looks up the position of the images in the input sequence and inserts the image embeddings in their respective positions.

the set of additional parameters $\hat{\theta}$ and the final augmented hidden dimension $\hat{d}_{llm}$. The difference $\Delta\hat{d} = \hat{d}_{llm} - d_{llm} > 0$ measures the amount of parameters augmentation.

3.3 Architecture

3.3.1 Prompting

To aid the LLM to perform multi-modal co-referencing, we introduce special tokens to delineate the beginning and end of dialogues, as well as the beginning and end of images. We also introduce a special token ($< im >$) to mark the positions of images in the text. This will then be used by the Multi-Modal Interleaver shown in Figure 2 to insert the image embeddings into the text embeddings at the same position the image was in the input, i.e, fusing the special token embeddings with the respective image embeddings.

```
<dialogue>
    ...
    <image><im></image>
    ...
</dialogue>
```

3.3.2 Linear Layer

Image embeddings are obtained from a frozen image encoder v_ϕ that maps the collection of p images to vectors $\mathbf{x}^{v_m} \in \mathbb{R}^{p \times d_{v_m}}$ (e.g., CLIP (Radford et al., 2021)). These visual embeddings need to be aligned with text embeddings coming from the LLM $h_\theta(x^t) \in \mathbb{R}^{d_{llm}}$ (which also includes the placeholder $< im >$ tokens).

To achieve this, we simply apply a linear transformation by multipling with $\mathsf{W}_l \in \mathbb{R}^{d_{v_m} \times d_{llm}}$ similar to (Lyu et al., 2023; Koh et al., 2023b):

$$\mathbf{x}^v = W_l^T \mathbf{x}^{v_m} \quad , \quad \mathbf{x}^v \in \mathbb{R}^{p \times d_{llm}} \tag{1}$$

3.3.3 Cross-modality Normalization

Previous works have demonstrated neural architectures to be especially sensitive to the statistics of their activations, exemplified by popular layer normalization blocks such as LayerNorm or RMS (Zhang & Sennrich, 2019) used in LLM architectures. As we wish to fuse the image embeddings $\mathbf{x}^v$ onto LLM token representations $\mathbf{x}^t$, we compute the magnitude of $\mathbf{x}^t$, averaged across the interleaved sequence, and use them to rescale $\mathbf{x}^v$ component-wise. More precisely, considering a sequence of n textual embeddings $\mathbf{x}^t \in \mathbf{R}^{n \times d_{llm}}$:

$$\sigma^t = \sqrt{\frac{1}{d_{llm} \times n} \sum_{i,j} \left(\mathbf{x}^t_{ij} - \mu(\mathbf{x}^t)\right)^2} \tag{2}$$

$$\mathbf{x}^v \leftarrow \mathbf{x}^v / \sigma^t \tag{3}$$

with $0 < i < n$ indexing the tokens sequence, $0 < j < d_{llm}$ indexing the features and $\mu(\cdot)$ the mean over both sequence and feature dimensions.

3.3.4 Multi-Modal Interleaver

The role of the Multi-Modal Interleaver (shown in Figure 2, right-hand side) is to preserve the integrity of the sequence of multi-modal input. Since the images are separated from the text so that they can be processed by the image encoder, it is possible to lose the original order of the multi-modal input. Recent works (Lyu et al., 2023) concatenate the multi-modal aligned embeddings, but this changes the sequence of the inputs that will be processed by the model. We replace the removed images with the special token $< im >$ which marks the position of the images. These special tokens will be replaced with cross-modalities normalized embeddings by the Multi-Modal Interleaver.

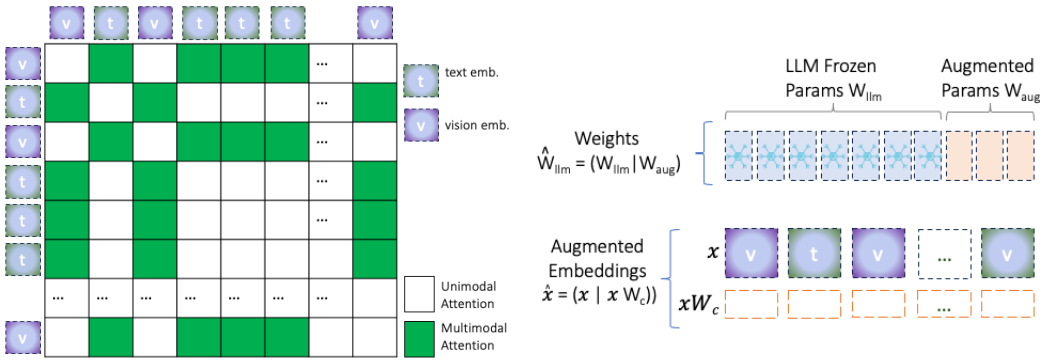

Figure 3:
LEFT: Multi-modal Attention has the advantage of implying both self and cross attention using the same parameters, while preserving the original order of the interleaved image-text sequence.
RIGHT: LLM parameters are depicted with the ice icon, showing they are frozen. The Parameter Augmented LLM weights are obtained by concatenating the frozen weights of the LLM and the augmented parameters.

We fuse the embeddings of the special token $< im >$ with the respective aligned image embeddings by a simple elementwise addition operation. The resulting multi-modal embeddings are then passed on to the rest of the LLM Layers as interleaved text and image embeddings as seen in Figure 2. This has the advantage of performing the essential cross attention operation as shown in Figure 3a without the use of a separate module. It also preserves the distances between tokens and images in the dialogue, which is likely helpful for co-reference resolution.

3.3.5 PARAMETER AUGMENTATION

LLMs have been shown to exhibit the catastrophic forgetting phenomena after being fine-tuned on data with a different underlying distribution (Zhai et al., 2023). A straightforward mitigation is to freeze the LLM (Zhai et al., 2023). This is the foundation principle behind Parameter Augmentation i.e., we freeze the uni-modal LLM parameters θ and introduce separate parameters $\hat{\theta}$ to map separate modalities together as seen in Figures 2 and 3b. We argue that this preserves the robustness of LLMs, allowing the transfer of their high-quality representations to other modalities. To upcycle the LLM $h_\theta(\cdot)$ to $h_{\theta,\hat{\theta}}(\cdot)$, we augment the modules at each layer by extending the hidden dimension d_{llm} through concatenation of additional weights: for each existing LLM weight matrix $W_{llm} \in \mathbf{R}^{r \times d_{llm}}$, we create $\hat{W}_{llm} = (W_{llm}|W_{aug})$, with trainable weights where $W_{aug} \in \mathbf{R}^{r \times \Delta\hat{d}}$. All subsequent operations (attention, normalization, feed forward) are therefore between inputs and augmented weights $\hat{W}_{llm}$. We demonstrate that even a small increase $\Delta\hat{d}$ along the hidden dimension is sufficient for the augmented LLM to learn complex relationships such as those in Figure 1. By only optimizing W_{aug} (and freezing W_{llm}), our approach reaps computation and memory benefits. As a comparison, the baseline (Awadalla et al., 2023) increases the LLM parameters by 18.7% while our approach increases it by 5.3%.

After the multi-model interleaver, the sequence of n fused image and token vectors $\mathbf{x} \in \mathbb{R}^{n \times d_{llm}}$ are still of the dimension of the original LLM d_{llm}. To map them to $\hat{\mathbf{x}} \in \mathbb{R}^{n \times \hat{d}_{llm}}$, we add an additional linear adapter $W_c \in \mathbb{R}^{d_{llm} \times \Delta\hat{d}}$:

$$\hat{\mathbf{x}} = (\mathbf{x}|\mathbf{x}W_c) \quad , \quad \hat{\mathbf{x}} \in \mathbb{R}^{n \times \hat{d}_{llm}} \tag{4}$$

By concatenating the augmented dimensions with the original embedding themselves, we hope to keep intact the spatial information encoded in pre-trained LLM embeddings (also see illustration in Figure 3b).

We can now optimize the negative log likelihood $\mathcal{L}_{\hat{\theta}}$ of the augmented LLM, with respect to $\hat{\theta}$. Element x_i at any position i below can be either image or text, their order determined by the interleaved sequence:

$$\mathcal{L}_{\hat{\theta}} = -\frac{1}{B}\sum_{j=0}^{B}\sum_{i=1}^{n} \log\Big(h_{\theta,\hat{\theta}}(x_i|x_0, \cdots, x_{i-1})\Big) \tag{5}$$

with j indexing the dataset of size B.

4 EXPERIMENT SET UP

We experiment on an image selection (Wanigasekara et al., 2022; 2023) and a specially curated image retrieval dataset adapted from (Saha et al., 2018). We measure performance for image selection using accuracy while we use classification metrics (accuracy, precision, recall, F1) for the image retrieval task. For both datasets, we fine-tuned the models for only 1 epoch.

4.1 DATASETS

Table 3 shows the statistics of the curated MMDA dataset. The average number of tokens after prompting is $717.5(\pm 413.6)$, $713.7(\pm 416.9)$ and $707.9(\pm 409.2)$ for the train, validation and test sets respectively. The average number of images per dialogue is approximately $32(\pm 20)$ for all data splits. The average number of utterances per dialogue is approximately $13(\pm 7)$ for all data splits. The ratio of positive to negatively (P:N) annotated images at the final utterance is $1:6$ for all data splits. A image is considered positive if it is relevant to the user's query that involves co-reference resolution.

The Multi-Modal Context Carryover (MMCC) dataset (Wanigasekara et al., 2022; 2023) is similar to datasets used in VQA and Image Captioning i.e. images can be logically separated from text. The Multi-Modal Domain Aware (MMDA) (Saha et al., 2018) contains an average of 32 images per dialogue, logically separating the images from text can impede performance. Also, in the MMDA dataset multiple images can be correct while in the MMCC dataset, there is only one correct image.

	Train	Valid	Test
# Dialogues	38,843	8,373	8,478
Avg # Tokens	717.5	713.7	707.9
Avg # Images	32.2	32.2	31.9
Avg # Utterances	13.3	13.1	13.1
Ratio P:N	1:6	1:6	1:6

Table 1: Dataset Statistics for the curated MMDA dataset. The ratio P:N is the ratio of the positively annotated images against negatively annotated images at the terminal utterance. A label is considered positive if it is relevant to the user's query that involves co-reference resolution.

The **Image Selection** task is performed on the Multi-Modal Context Carryover (MMCC) dataset (Wanigasekara et al., 2022; 2023). This dataset has $33k$ entries containing 3 product images, their descriptions and a selection criteria. Given the list of products images, product descriptions and selection criteria, the task is to select the product which has the highest probability to match the criteria. We model this as a generation rather than a classification task, where the LLM generates the index of the product image. We prompt this as shown below for both OpenFlamingo (Awadalla et al., 2023) and the Parameter Augmented LLM approach:

```
Image <position><image><im><image>
<description>
Action: Given the list of images,
determine the position of the image
that satisfies the criteria
Criteria: <criteria>
Position: <MASK>
```

The **Image Retrieval** task is performed on the Multi-Modal Domain-Aware (MMDA) (Saha et al., 2018) dataset. We require that contexts have at least 1 multi-modal utterance and that the last utterance (where inference happens) have both positive and negative labelled data. We discard all dialogues that do not meet this criteria. During inference, we shuffle the list of positive and negative images and predict whether each one belongs to the last utterance or not.

```
User: ...
System: ...
Question: Is <image><im></image>
a good match?
Answer: <MASK>
...
```

4.2 Pretrained vision encoders and multi-modal LLMs

Pretrained vision encoders: We are able to directly use pretrained vision encoders like CLIP and BLIP as simple baselines for the image selection task in a zero-shot manner. We extract product image and product description text embeddings separately. The image with the highest cosine similarity with the textual referring utterance is chosen as the selected image. For the image retrieval task in a dialogue setting, the dialogue contexts are too long for the direct use of CLIP and BLIP encoders (717 ± 410 tokens) so we use OpenFlamingo as our baseline.

Pretrained multi-modal LLMs: OpenFlamingo (Awadalla et al., 2023) is the publicly available version of the Flamingo (Alayrac et al., 2022) LLM. The 9B variant of OpenFlamingo is made up of a 7B MPT LLM (Team, 2023) with CLIP as the image encoder. It is pretrained on the LAION multi-modal dataset and so has some zero-shot capabilities. In the image selection task, we prompt the model to generate the index of the relevant image. In the image retrieval task, we prompt the model to generate Yes/No for each image in the candidate solution pool.

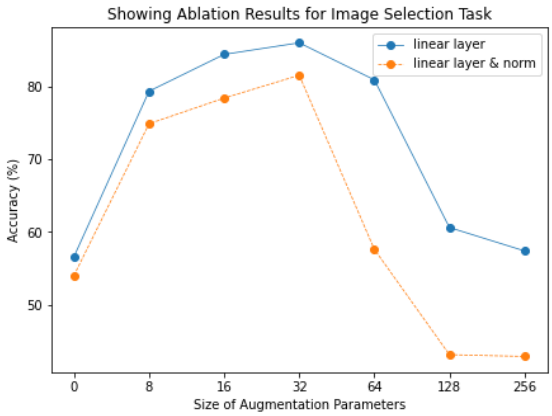

Figure 4: Ablation results of Parameter-Augmented LLM on image selection task, showing accuracy as a function of $\Delta\hat{d}$.

4.3 Augmented LLM

The parameter augmentation technique we propose is LLM-agnostic. In our experiments, we use the Open LLaMA (Touvron et al., 2023a) 7B version 1 from huggingface repository and make the parameter augmentation changes. This model has a hidden dimension size of 4096, we explore augmentations ranging from $\Delta\hat{d} = 0$ to $\Delta\hat{d} = 256$.

We prompt our Parameter Augmented LLaMA in a similar fashion as OpenFlamingo and perform ablation experiments for both image selection and retrieval tasks, treating $\Delta\hat{d}$ as a hyperparameter. Our Parameter Augmented LLaMA model is at a disadvantage when compared to the OpenFlamingo model because the OpenFlamingo model has been further fine-tuned on multi-modal tasks using 2B image-text pairs from the LAION (Schuhmann et al., 2022) dataset. Thus, our Parameter Augmented LLaMA model does not have as good zero-shot or in-context learning capabilities as OpenFlamingo.

5 Results

The multi-modal multi-turn setting adds complexity to the co-referencing problem since each user utterance can reference *any* system utterance in the previous turns as seen in appendix Figures 5 and 6. The dialogues are also long for the MMDA dataset, i.e. average of 717 tokens compared to an average of 82 tokens for the MMCC dataset. In this paper, we use image retrieval metrics as a proxy to measure multi-modal co-referencing.

5.1 Image Selection Results

Figure 4 shows ablation experiment results on the image selection dataset (Wanigasekara et al., 2022; 2023). The "linear layer" only includes the linear module shown in Figure 2 and the "linear layer & norm" has the linear module and the cross RMS norm module. We sweep the hyperparameter $\Delta\hat{d}$ from 0 to 256 where 0 indicates no augmentation. We observe parameter augmentation range $8 - 64$ to be the best setting for both "linear layer" and "linear layer & norm".

Table 2 shows the results using different family of models on an image selection dataset. We obtained the LSTM results from previous SOTA (Wanigasekara et al., 2022) for the image selection task. For the image encoders, we observe that CLIP (Radford et al., 2021) has better performance compared to BLIP (Li et al., 2022). Prompting and fine-tuning OpenFlamingo resulted in the best performance overall. Parameter Augmented LLaMA with $\Delta\hat{d} = 32$ performed better than OpenFlamingo in zero-shot but was outperformed in in-context and fine-tuned settings. OpenFlamingo (Awadalla et al., 2023) was pretrained on $2B$ image-text pairs from LAION (Schuhmann et al., 2022) so the comparison is not fair but this still illustrates how a uni-modal LLM such as LLaMA can transfer its capabilities to the multi-modal setting through parameters augmentation.

Model	Set Up	Accuracy
BLIP	zero-shot	44.17%
CLIP	zero-shot	77.40%
LSTM + CLIP	Fine-Tuning	84.84%
LSTM + ALBEF	Fine-Tuning	86.17%
OpenFlamingo	zero-shot	32.40%
	In Context	38.48%
	Fine-Tuning	**90.12%**
Parameter Augmented LLaMA $\Delta\hat{d} = 32$	zero-shot	34.51%
	In Context	34.92%
	Fine-Tuning	85.95%

Table 2: Showing results of image-text models, ensemble, OpenFlamingo and Parameter-Augmented LLM on image selection task. LSTM results are from previous state of the art (Wanigasekara et al., 2022)

.

Model	Experiment Set Up	Accuracy	Precision	Recall	F1
OpenFlamingo	zero-shot	35.67%	0.3472	0.9635	0.4621
	In-Context	36.14%	0.3486	**0.9651**	0.4648
	Fine-Tuning	76.70%	0.6953	0.8235	0.7240
Parameter Augmented LLaMA Fine-Tuning	$\Delta\hat{d} = 0$	66.77%	0.5583	0.8334	0.5995
	Norm & $\Delta\hat{d} = 0$	46.40%	0.3950	0.9279	0.4915
	$\Delta\hat{d} = 64$	77.89%	**0.7118**	0.9334	**0.7727**
	Norm & $\Delta\hat{d} = 64$	70.93%	0.6519	0.9096	0.7135
	$\Delta\hat{d} = 128$	77.84%	0.6702	0.6510	0.6228
	Norm & $\Delta\hat{d} = 128$	70.64%	0.4846	0.3058	0.3438
	$\Delta\hat{d} = 256$	78.73%	0.6373	0.5090	0.5323
	Norm & $\Delta\hat{d} = 256$	**79.05%**	0.6595	0.8437	0.7122

Table 3: Showing results of a OpenFlamingo and our Parameter-Augmented LLM for image retrieval task applied on the MMDA dataset.

5.2 Image Retrieval Results

In Table 3, we show the performance of multi-modal LLMs on the image retrieval dataset. We observe poor zero-shot and in-context performance for both datasets (Tables 2, 3) using Open Flamingo, highlighting the difference between our datasets and image-text datasets which are used in pretraining. After fine-tuning, the Parameter Augmented LLaMA ($\Delta\hat{d} = 64$) outperforms fine-tuned OpenFlamingo for precision ($+0.016$), recall ($+0.1099$) and F1 ($+0.0487$) while $\Delta\hat{d} = 256$ with Cross Normalization has $+2.35\%$ better accuracy (Table 3).

6 Conclusion

By crossing both dialog turns and modalities, multi-modal co-reference resolution presents notorious challenges. In this paper, we explore the possibility to leverage existing pre-trained LLM capabilities and offer a simple and robust parameter augmentation technique that does not require additional multi-modal pre-training tasks. We demonstrate competitive results in image selection and best results in the image retrieval dataset compared to a baseline pre-trained on billions of multi-modal examples. This opens up the possibility to leverage the flurry of recently released open-source models, both by the NLP and vision communities, to improve upon multi-modal setups.

REFERENCES

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

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

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. Gravl-bert: graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 285–297, 2022.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10877, 2023.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models. *arXiv preprint arXiv:2305.17216*, 2023a.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023b.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*, 2021.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, et al. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 813–830, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

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

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

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023.

OpenAI. Gpt-4 technical report, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

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

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023.

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

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

Prashan Wanigasekara, Nalin Gupta, Fan Yang, Emre Barut, Zeynab Raeesy, Kechen Qin, Stephen Rawls, Xinyue Liu, Chengwei Su, and Spurthi Sandiri. Multimodal context carryover. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 417–428, 2022.

Prashan Wanigasekara, Rafid Al-Humaimidi, Turan Gojayev, Niloofar Gheissari, Achal Dave, Stephen Rawls, Fan Yang, Kechen Qin, Nalin Gupta, Spurthi Sandiri, et al. Visual item selection with voice assistants: A systems perspective. In *Companion Proceedings of the ACM Web Conference 2023*, pp. 500–507, 2023.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*, 2021.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

A APPENDIX

Figure 5 and 6 show sample result of Parameter Augmented LLaMA on the image retrieval MMDA dataset with F1 score of 1.0 and 0.8 respectively. A red box around an image refers to a negatively labelled image while a green box refers to a positively labelled image. The yellow box refers to the image the user is currently using as an example. The models then predict Yes/No given a list of images.

Our approach is able to differentiate between styles of similar images as we show in the appendix Figure 5 where the candidate products are both sandles but different styles, this is more granular than differentiating unrelated objects e.g., sandles vs chair. In Figure 6, we see similar capabilities over more utterances. We attribute the false negative result (prediction No but the box is green) in Figure 6 as a mis-annotation because the shoe is not similar to co-referenced shoe (shoe with yellow border) and is not made of strap material nor a high top as specified by user utterance.

In the OpenFlamingo architecture, $1.3B$ of the $9B$ parameters are optimized, this accounts for 18.6% with respect to its uni-modal LLM ($7B$). In the parameter augmented setting with $\Delta\hat{d} = 64$, we optimize $370M$ parameters (5.3% of uni-modal LLM) which is a more resource efficient setup. Thus, our model optimizes 13.3% fewer parameters with respect to the uni-modal LLM (in both cases, the uni-modal LLM is 7B).

With only one epoch of fine-tuning, our Parameter Augmented LLM approach, which was not initially pre-trained on a multi-modal dataset, was able to achieve competitive results compared to OpenFlamingo. This shows the possibility of transferring benefits of uni-modality to multi-modality using parameter augmentation.

In the image selection results in Figure 4, we see a significant drop in performance for $\Delta\hat{d} = 128$ and $\Delta\hat{d} = 256$. This is because it introduces more than $1.5\times$ more parameters compared to the other augmentations. The image selection dataset is comparatively smaller and has a total of approximately $3M$ tokens, and training on one epoch is insufficient for the higher number of parameters. For image retrieval, the dataset is comparatively larger and has approximately $30M$ tokens, and so we see steady performance improvements with higher augmentations.

The augmented LLM variant also resulted in the best performance for the image retrieval dataset, exceeding that of a $1.2\times$ model, trained on $20k\times$ more data while optimizing $3.5\times$ fewer parameters. This is in line with our hypothesis - we reap more benefits from using parameter augmentation when the co-reference task is difficult: we see more gains for the MMDA dataset than the MMCC dataset where the co-reference is simpler.

For the image selection task based on Figure 4, for augmentation 128 and 256, the Cross Normalization is vastly outperformed by the normalization ablation. Overall, variants with cross normalization are outperformed by the variants without normalization. We observe a different trend for image retrieval in that the $\Delta\hat{d} = 256$ augmented LLaMA, with normalization performing better than without normalization, setting the best accuracy result (see Table 3). We conclude that the Cross Normalization is more beneficial in high augmentation settings and in more complex task of image retrieval (where co-referencing happens between multiple dialogue turns and multiple images).

Figure 5: Sample Image Retrieval result for Parameter Augmented LLaMA on a dialogue with 5 utterances and 5 images before the final utterance. In this case, F1=1.0. The green box indicates that the image is relevant, and the red box indicates that the image is not relevant to the user query.

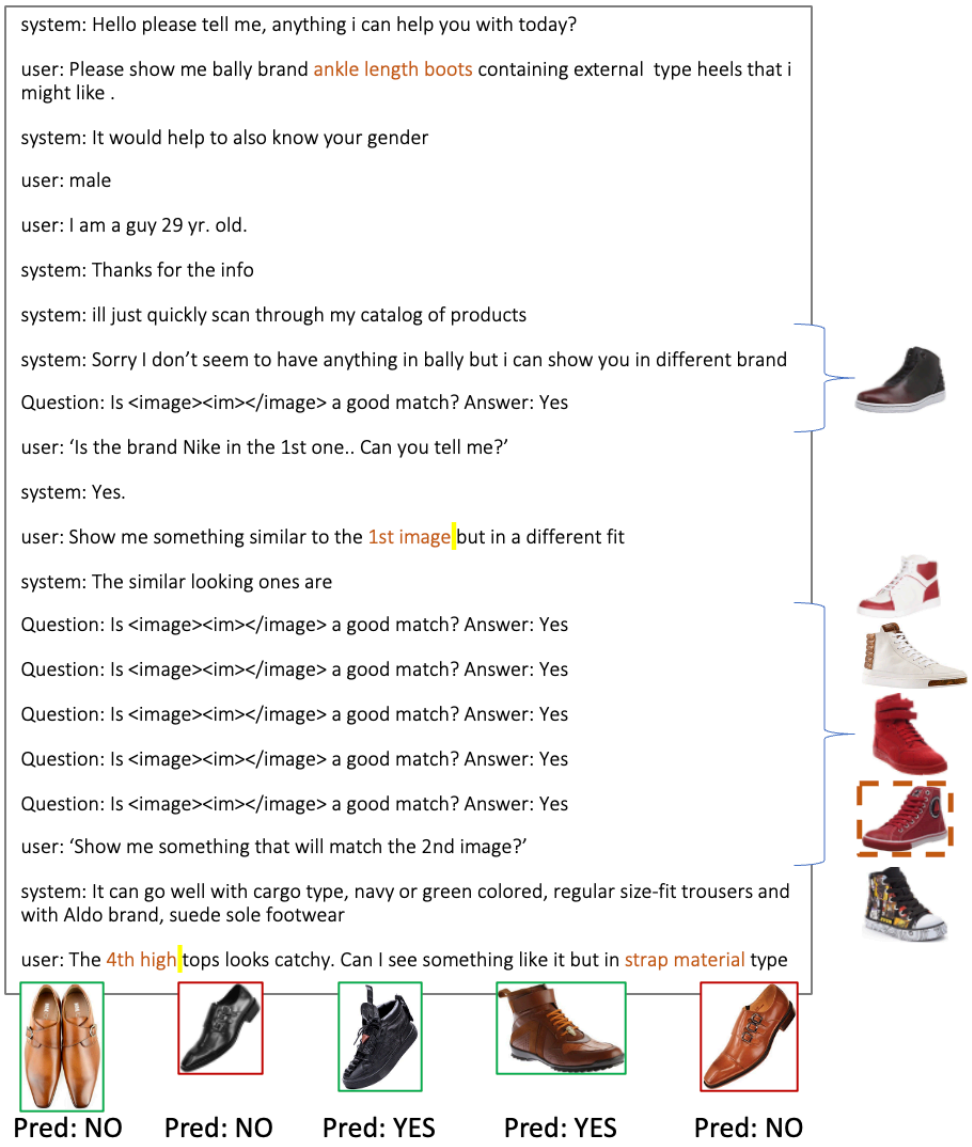

Figure 6: Sample Image Retrieval result for Parameter Augmented LLaMA on a dialogue with 16 utterances and 6 images before the final utterance. In this case, F1=0.8. The green box indicates that the image is relevant, and the red box indicates that the image is not relevant to the user query. The orange dotted box refers to the image the user is currently using as an example.