Towards Multi-Modal Co-Reference Resolution in Conversational Shopping Agents

Samuel Osebe*[1], Prashan Wanigasekara*[2], Thomas Gueudre[2], Thanh Tran[2],
Rahul Sharma[2], Fan Yang[2], Qian Hu[2], Weitong Ruan[2], Emre Barut[2], Chengwei Su[2]
[1]University of Massachusetts Amherst, [2] Amazon AGI Foundations
sosebe@umass.edu,
{wprasha, tgueudre, tdt, zarahuls, fyaamz, huqia, weiton, ebarut, chengwes}@amazon.com

Abstract

The context of modern smart voice assistants is 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 human in a shopping use case. We propose 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 cross-attention baseline while reducing the number of parameters being trained by $4\times$.

Keywords: multi-modality, co-referencing, parameter-augmentation

1. Introduction

Recent advancements in multi-modal large language models (MLLMS) have pushed the capabilities of conversational agents, extending beyond processing and generating human-like text to include understanding and integrating multiple modalities such as images and audio. These advancements have led to substantial progress in tasks like image captioning (Lin et al., 2014), image classification (Russakovsky et al., 2015) and visual question answering (Goyal et al., 2017). However, these tasks often have a clear division between the text and images, not fully reflecting the complex, interwoven nature of the inputs encountered by conversational agents. This intertwining of visual and textual inputs is more pronounced in the environments like online shopping, where users seamlessly shift between textual and visual references.

Addressing this gap, Multi-modal Co-reference Resolution (MCR) emerges as a critical challenge, aiming to connect language and visual content by mapping textual references to their corresponding spatial regions in images. In this work we focus on MCR within the context of conversational agents in the shopping domain where the challenge is amplified by the vast diversity of products and the ambiguity of natural language descriptions, marking a stark contrast to areas like visual question answering (VQA) and image captioning. Efforts to address MCR for conversational agents have been relatively limited, further compounded by the fact that most multi-modal dialogue datasets (Zang et al., 2021; Kottur et al., 2021), contain very few images among the dialogues. In contrast, a typical dialogue in the shopping context can involve $5-66$ utterances, with an average of 32 images, highlighting the need for specialized attention to MCR in this domain. Figure 1 shows a sample dialogue for the shopping use case.

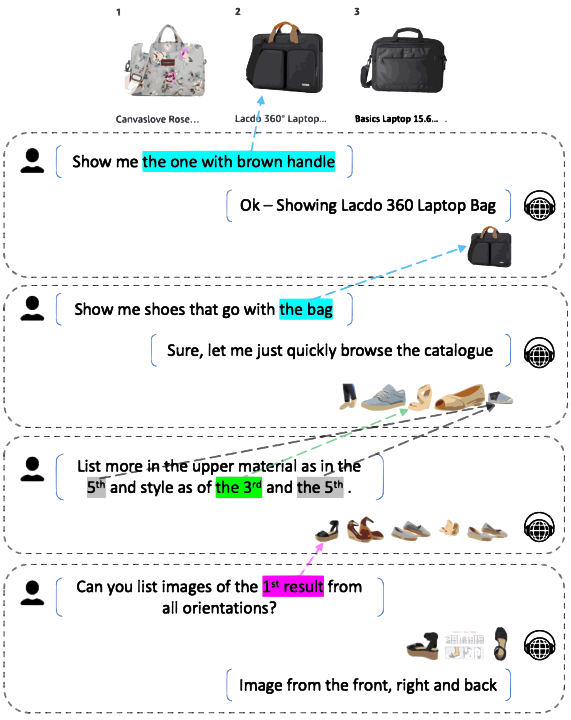

Figure 1: An example of a multi-turn dialogue with multi-modal co-referencing. The co-references are color coded and shown by arrows.

*equal contribution

The Seventh Workshop on e-Commerce and NLP (ECNLP 7), pages 8–18

Among the recent efforts (Lee et al., 2022; Guo et al., 2022; Chen et al., 2023) to address MCR for conversational agents, an encompassing strategy has been the end-to-end training of multi-modal transformer architectures. While effective, this strategy demands significant computational resources, manifesting in both a high number of parameters and extensive training time. To mitigate these challenges, we propose a novel technique that leverages existing unimodal large language models (LLMs) and adapt them for multi-modal inputs and outputs. Our approach augments the weights of pre-trained unimodal LLMs to learn an alignment with the pre-trained visual encoder's embeddings, thereby converting them into multi-modal system. This method significantly reduces the number of parameters to be trained along with a notable improvement in the MCR performance.

To evaluate the effectiveness of our methodology in practical conversational tasks within the shopping domain, we focus on two key areas: i) image selection and ii) image retrieval. The image selection task leverages textual and visual attributes to identify and select the most relevant product from a list of options; e.g., the utterance "show me the bag with brown handles" in Figure 1. For this, we employ the Multi-Modal Context Carryover (MMCC) (Wanigasekara et al., 2022) dataset to assess our model's performance in accurately selecting the correct product based on the given criteria. In the image retrieval task, the objective is to identify relevant products at the final turn of a multi-turn, multi-modal dialogue.

We use the Multi-Modal Domain Aware (MMDA) dataset (Saha et al., 2018), which is rich in image-inclusive dialogues, for evaluating our model's performance for an image retrieval task. Our results demonstrate a significant improvement over existing models, including the pretrained multi-modal cross-attention model, OpenFlamingo (Awadalla et al., 2023). We achieve an increase of approximately 5 points in F1 score, while training 4x fewer parameters. This underscores the efficiency of our proposed parameter augmentation methodology for 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). The goal of this work is aligning the embeddings of such visual models with embeddings of pretrained language models efficiently. Alignment can be classified as either natural language alignment or embedding alignment.

2.1. Multi-Modal Alignment Approaches

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. This is 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(Alayrac et al., 2022), Open-Flamingo (Awadalla et al., 2023) and BLIP (Li et al., 2022) use cross attention and contrastive learning objectives. BLIP2 (Li et al., 2023) proposes a querying transformer to learn queries for the visual embeddings. Mini-GPT4 (Zhu et al., 2023) proposes to only train a linear projection layer to project visual embeddings to text space. 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) for a similar projection. Most recently, GILL(Koh et al., 2023a) uses a linear projection and a learnable query embeddings module.

Currently. there are closed-source pipelines such as GPT-4 (OpenAI, 2023; Yang et al., 2023) and GEMINI (Team 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 using a multi-modal mixture of experts setup, we do not compare our work with them.

2.2. Multi-Modal Co-Reference Resolution

MCR bridges the gap between language and images by mapping the text to spatial regions being referred. A closely related field, Visual Grounding (VG) seeks to align text queries with their corresponding locations in images. In the VG domain, JR-net (Jain and Gandhi, 2022) is one of the SOTA methods; it separately encodes images and queries and then employs a sophisticated joint reasoning and fusion method to generate results. VLT (Ding et al., 2023) is another method that transforms the image data into the same space as language token embeddings and uses a masked decoder to

locate targets. Several datasets have also been introduced to support the research in the direction of MCR which includes CIN (Goel et al., 2022) which is rich in co-reference chains and grounding annotations, and others (Parcalabescu et al., 2022; Ramanathan et al., 2014; Cui et al., 2021; Hong et al., 2023) that link textual mentions of people with their images.

More recently, SIMMC2.0 and SIMMC2.1 datasets are introduced as challenges in DSTC. These datasets encompass 11,000 task-oriented dialogues for shopping scenarios with photorealistic scenes, spurring the development of numerous multi-modal methods tailored to conversational agents. In (Lee et al., 2022) proposes a multi-modal encoder-decoder model that offers a unified solution for various tasks associated with situated conversational agents, including MCR. GraVL (Guo et al., 2022) introduce an innovative approach to merge Graph Neural Networks with VL BERT capturing visual relationships alongside dialogue and metadata for nuanced understanding. SHIKRA (Chen et al., 2023) stands out by proposing a multi-modal model capable of engaging in referential dialogue, enabling users to input specific image regions and responding by referencing the pertinent areas if required.

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 based on modalities, even though the model may accept interleaved multi-modal inputs. 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 observe poor performance when performing an image retrieval task over multiple co-referenced images. We test this hypothesis using our proposed approach, which preserves the sequence of the multi-modal information during processing. We use OpenFlamingo as baseline for our experiments.

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 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 $\mathrm{W}_l \in \mathbb{R}^{d_{v_m} \times d_{llm}}$ similar to (Lyu et al., 2023; Koh et al., 2023b):

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

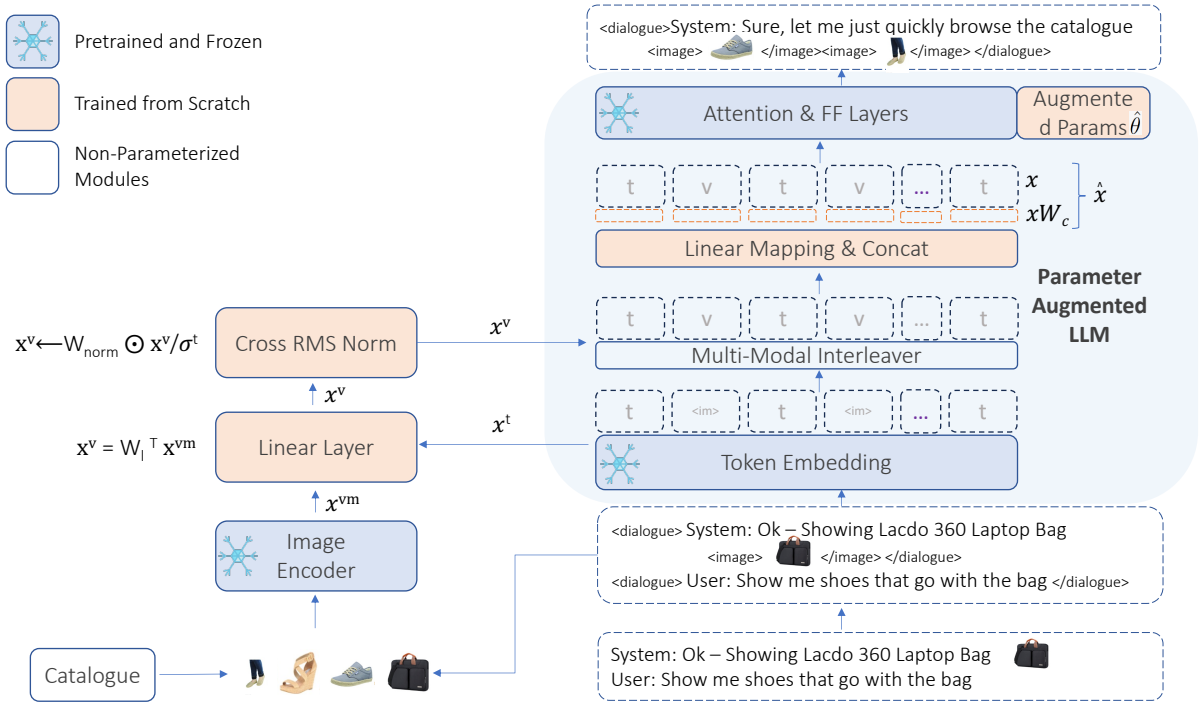

Figure 2: LLM-Agnostic Architecture for Parameter Augmentation. Boxes with t and v refers to text and image embeddings, respectively. We optimize the linear layer, cross RMS normalization module, and the augmented parameters, the rest of the LLM remains frozen. The Multi-Modal Interleaver looks up the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

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 and Sennrich, 2019) used in LLM architectures. This problem is accentuated in a multi-modal setup; indeed, differences in activations distributions for visual and textual inputs require different normalizations (BatchNorm in Vision vs LayerNorm in NLP) (Shen et al., 2020). As we wish to fuse both 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 $\mathbf{x}^t \in \mathbf{R}^{n \times d_{llm}}$:

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

$$\mathbf{x}^v \leftarrow \mathbf{x}^v / \sigma^t, \quad (3)$$

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. 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 3 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.

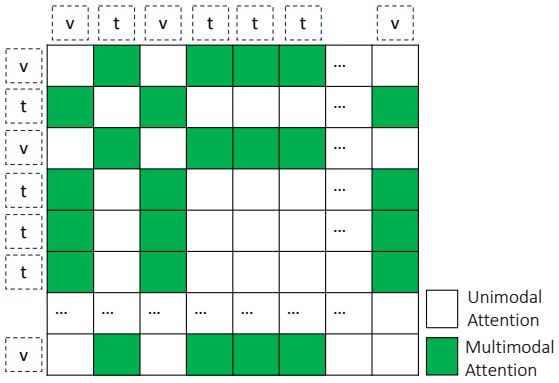

Figure 3: Multi-modal Attention has the advantage of implying both self and cross attention using the same parameters, while preserving the original order of the interleaved image-text sequence.

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 4. We argue that this preserves the robustness of LLMs, allowing the transfer of their high-quality representations to other modalities. To upcycle the LLM $h_\theta(\cdot)$

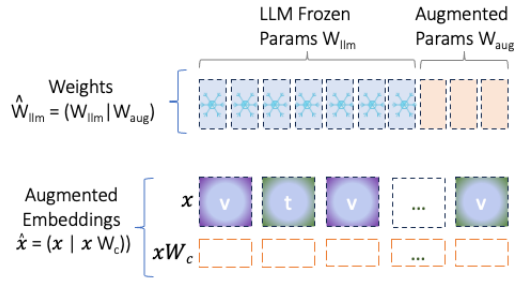

Figure 4: LLM parameters are depicted with the ice icon, showing they are frozen. The Parameter Augmented LLM weights are obtained by concatenating the frozen weights of the LLM and the augmented parameters.

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 pretrained LLM embeddings (also see illustration in Figure 4).

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:

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

where j indexes 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-tune the models for only 1 epoch.

4.1. Datasets

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, unlike the MMCC dataset which has only one correct image.

The **Image Selection** task is performed on the Multi-Modal Context Carryover (MMCC) dataset (Wanigasekara et al., 2022, 2023). This dataset

	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.

has $33k$ entries, each containing 3 product images, their descriptions and 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:

```
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 while for the image retrieval task, we prompt the model to generate a binary answer (Yes / No) for each image in the candidates.

4.3. Augmented LLM

The parameter augmentation technique we propose is LLM-agnostic. In our experiments, we augment the parameters of the Open LLaMA (Touvron et al., 2023a) 7B model. This model has a hidden dimension size of 4096, we explore augmentations in the range $\Delta\hat{d} = 0$ to $\Delta\hat{d} = 256$.

We prompt the augmented LLaMA similarly as OpenFlamingo and perform ablation experiments for both image selection and retrieval tasks, treating $\Delta\hat{d}$ as a hyperparameter. This 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, augmented LLaMA is not directly comparable with OpenFlamingo in zero-shot or in-context learning and is disadvantaged for fine-tuning. However, we see that it out-performs pretrained OpenFlamingo as shown in table 3

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 Figures 6 and 7. In this paper, we use image retrieval metrics as a proxy to measure multi-modal co-referencing.

5.1. Image Selection Results

Figure 5 shows ablation experiment results on the MMCC 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

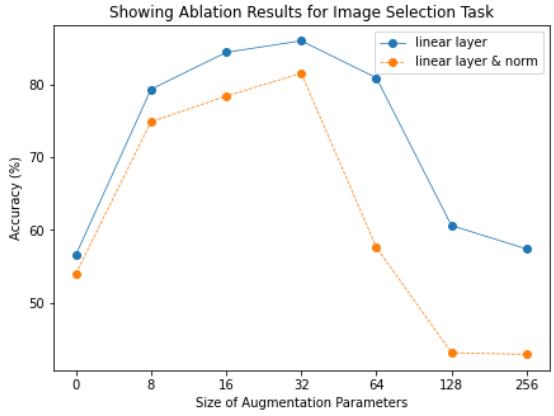

Figure 5: Ablation results of Parameter-Augmented LLM on image selection task, showing accuracy as a function of $\Delta\hat{d}$.

Figure 6: 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.

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 Visual Language models on MMCC dataset. We obtain 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. We see that a uni-modal LLM such as LLaMA can transfer its capabilities to the multi-modal setting through parameter augmentation and approach the performance of a pretrained

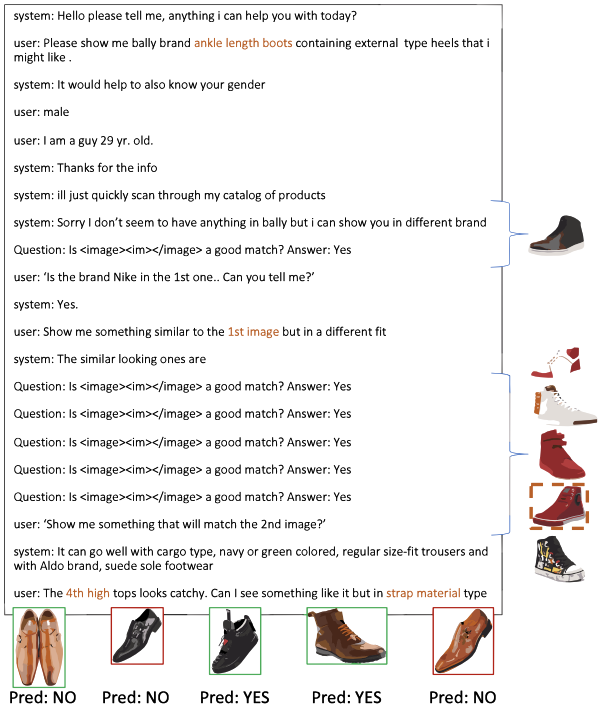

Figure 7: Sample Image Retrieval result for Parameter Augmented LLaMA on a dialogue with 16 utterances and 6 images before the final utterance. In this case, F1=0.8. The green box indicates that the image is relevant, and the red box indicates that the image is not relevant to the user query. The orange dashed box refers to the image the user is currently using as an example.

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.

5.2. Image Retrieval Results

In Table 3, we show the performance of multi-modal LLMs on the MMDA dataset. We observe poor zero-shot and in-context performance using Open

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
Linear Layer	Fine-Tuning	66.77%	0.5583	0.8334	0.5995
Parameter Augmented LLaMA Fine-Tuning	Linear Layer $\Delta\hat{d} = 64$	77.89%	**0.7118**	**0.9334**	**0.7727**
	Linear Layer & Norm $\Delta\hat{d} = 64$	70.93%	0.6519	0.9096	0.7135
	Linear Layer $\Delta\hat{d} = 128$	77.84%	0.6702	0.6510	0.6228
	Linear Layer & Norm $\Delta\hat{d} = 128$	70.64%	0.5997	0.9213	0.6851
	Linear Layer $\Delta\hat{d} = 256$	78.73%	0.6373	0.5090	0.5323
	Linear Layer & 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.

Flamingo, highlighting the difficulty of the task. After fine-tuning, the Parameter Augmented LLaMA ($\Delta\hat{d} = 64$) outperforms fine-tuned OpenFlamingo. This highlights the robustness of parameter augmentation over cross-attention.

5.3. Qualitative Analysis

Figure 6 and 7 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 dashed 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 Figure 6 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 7, we see similar capabilities over more utterances. We attribute the false negative result (prediction No but the box is green) in Figure 7 as a mis-annotation because the shoe is not similar to co-referenced shoe (shoe with orange dashed 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).

In the image selection results in Figure 5, 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 given the higher number of parameters. For image retrieval results (Table 3), 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 model with $1.2\times$ parameters, trained on $20k\times$ more data while optimizing $3.5\times$ fewer parameters. We see more gains for the MMDA dataset than the MMCC dataset, where the co-reference is simpler. This is in line with our hypothesis - we reap more benefits from using parameter augmentation when the degree of multi-modal co-referencing increases.

For the image selection task based on Figure 5, for augmentation $\Delta\hat{d} = 128$ and $\Delta\hat{d} = 256$, the Cross Normalization is significantly 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). However, we observe more over-fitting to the data when normalization is used, creating the need for a better design for multi-modal normalization. We will explore this in our future work.

6. Conclusion

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 cross-attention baseline pre-trained on billions of multi-modal examples.

7. References

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

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

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

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. *arXiv preprint arXiv:1902.10547*.

Claire Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1374–1384.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2023. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916.

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

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2022. Who are you referring to? weakly supervised coreference resolution with multimodal grounding. *arXiv preprint arXiv:2211.14563*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. 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*, pages 10867–10877.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Kanishk Jain and Vineet Gandhi. 2022. Comprehensive multi-modal interactions for referring image segmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

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

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

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

Hung Le, Nancy F Chen, and Steven CH Hoi. 2022. Multimodal dialogue state tracking. *arXiv preprint arXiv:2206.07898*.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee,

et al. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830.

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

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Yunyao Li and Angeliki Lazaridou. 2022. Proceedings of the 2022 conference on empirical methods in natural language processing: Industry track. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

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

OpenAI. 2023. Gpt-4 technical report.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vignesh Ramanathan, Armand Joulin, Percy Liang, and Li Fei-Fei. 2014. Linking people in videos with "their" names using coreference resolution. In *Computer Vision – ECCV 2014*, pages 95–110, Cham. Springer International Publishing.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

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

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

Sheng Shen, Zhewei Yao, Amir Gholami, Michael Mahoney, and Kurt Keutzer. 2020. Powernorm: Rethinking batch normalization in transformers. In *International Conference on Machine Learning*, pages 8741–8751. PMLR.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

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

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

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

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

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

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

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

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

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

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

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

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