# Vision-Language Pre-trained Model for Visual Grounding with Reasoning Requirement

*William Sutanto*


Master of Science
School of Informatics
University of Edinburgh
2024

# Abstract

Multimodal AI represents a significant milestone in human-machine interaction. One of the fundamental task in Multimodal AI is Visual grounding (VG), a vision-languange task focusing on locating objects based on a given query expression. In simple VG, all the necessary information to locate the object is contained in the query through visual description or spatial information. In this project we focus on VG with reasoning requirement, where the model must comprehend a scene story first as knowledge to accurately understand the query.

We evaluate the performance of vision-language pre-trained model, specifically Kosmos-2 and OFA in solving the VG task. The evaluation is conducted in two settings: zero shot and fine-tuned. Additionally, we investigate the effect of query modification by leveraging Llama 3's reasoning capability. Our findings shows that Kosmos-2 and OFA has capability in VG, outperforming LeViLM, an existing model, in zero shot setting. We also discover that query modification with Llama 3 improves the accuracy in detecting bounding box. In fine-tuned setting, our fine-tuned OFA, a generalized model, achieves competitive results compared to LeViLM, a specialized model that built for VG.

# Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*William Sutanto*)

# Acknowledgements

Spending a year pursuing my Master's degree in Edinburgh has been one of the greatest experience of my life. I am grateful for the opportunity to meet remarkable and wonderful people and to live in this lovely city.

To Prof. Frank Keller—
Thank you for all the insight, guidance, and support throughout the project. I really appreciate your humbleness and approachability despite your extensive knowledge. Your egalitarian approach made it easy for me to collaborate in this project.

To LPDP / Indonesia Endowment Fund for Education—
Thank you for awarding me a scholarship to study abroad, at the University of Edinburgh. This opportunity has been a truly life-changing experience for me.

To my friends in Edinburgh—
Thank you for the fun, support, laughter, and our memorable trips together. Despite being far from home, you has made me feel at home.

Lastly, to my family—
Thank you for your unwavering support throughout the year.

# Table of Contents

# Chapter 1

# Introduction

## 1.1 Motivation

The advent of Multimodal AI has represented a significant landmark in human-machine interaction. With its capability to emulate human perception and understanding from multiple types of data simultaneously, including image, text, and audio, multimodal AI boosts the efficiency of AI systems and opens new possibilities that were unimaginable previously. For example, a robot equipped with multimodal understanding could navigate more naturally and navigate through complex environments [1], or improve personalization and safety in autonomous driving [2].

One of the fundamental tasks in multimodal AI is visual grounding (VG), illustrated in Figure 1.1 (left), a visual-language understanding to locate an object inside an image by bounding a box based on a given text phrase query. In a simple VG task, all the information needed to understand the object is available from the image and the short text query with simple vision-language alignment. It locates the position through visual appearance or spatial information. This is a fundamental task in multimodal AI, with wide potential for better human computer interaction [5], image retrieval, visual QA, and autonomous vehicle [6].

Song et al., [4] introduce a new task on VG that requires complex reasoning abilities, along with the new dataset, called Scene Knowledge Visual Grounding dataset (SKVG dataset). To identify the referenced object, the model must comprehend a narrative backstory provided alongside the image and the query. For example, in the given image illustrated in Figure 1.1 (right), the query is ”Alan’s dog”. Simple VG might understand the semantics of ”dog”, but this information is not enough since there are multiple dogs in the image. The system must identify Alan first from the scene knowledge to


Figure 1.1: **Left**: In basic VG on RefCOCO Dataset [3], the model identify the correct bounding box solely based on physical appearance mentioned in the query. **Right**: In VG with Scene Knowledge [4], the model must have reasoning ability to understand the given background story in order to identify the object provided in query. In this particular example, identification of Alan is needed.

determine which dog to choose.

In recent years, the ”pretrain-finetune” paradigm has demonstrated significant success across various domains, including in vision-language[7]. Following this approach, OFA (”One-for-All”)[8] introduced a modality-agnostic model that unifies vision and language representation. Pre-trained on multiple language and vision task, OFA achieved a remarkable performance in basic VG task after fine-tuning phase.

Another key advancement in multimodality is the emergence of Multimodal Large Language Models (MLLMs), such as LLaVA [9], Flamingo [10], BLIP-2 [11], Kosmos-2 [12], and GPT-4V[13]. These models have shown its advancement in understanding multimodal perception and have successfully widened the Large Language Model (LLM) potential to other tasks, such as image captioning, text-to-image generation, visual question answering, and also visual grounding. Notably, Kosmos-2 has shown its capability in basic visual grounding even in zero-shot setting.

## 1.2 Objective

Since VG with reasoning requirement is a crucial part in human machine interaction and pretrained visual language model has shown its potential in solving many multimodal task in zero or few shot settings, we intend to deeply investigate the possibility and performance of Kosmos-2 and OFA to solve this problem. The specific objective of this research is to evaluate Kosmos-2 and OFA capability on VG task in these specific settings: (1) zero-shot; (2) zero-shot with modified query generated by LLAMA-3 as additional reasoning pipeline; (3) Fine tuning with SK-VG dataset.

## 1.3 Timeliness and Novelty

The visual grounding with scene knowledge is a newly proposed task and dataset by [4]. To the best of our belief, there is no research yet on evaluating Kosmos-2 and OFA performance on this task type. Our proposed method of modifying the query by adding a reasoning pipeline with Large Language Model (LLM) before grounding could be the first. LLMs such as GPT[14, 15, 16], Vicuna[17], and Llama [18, 19, 20] have shown remarkable performances across a range of language tasks, including reading comprehension, question answering, common sense reasoning, code generation, natural language inference, and many more.

## 1.4 Thesis Structure

This dissertation is organized into six chapters. Chapter 2 reviews related works on visual grounding, previous approaches in solving SK with reasoning, as well as the explanation of Kosmos-2 and OFA models. Chapter 3 described the datasets used for training and evaluation. Chapter 4 proceeds with the methodology employed in evaluating performances. Chapter 5 presents the results along with a discussion. Finally, Chapter 6 draws conclusions and future work recommendations are offered.

# Chapter 2

# Related Work

## 2.1 Types of Visual Grounding

Visual Grounding(VG) aims to locate the region in the given image referred by the query expression in natural language. Various datasets and task formulations have been proposed to address the challenges of grounding.

### 2.1.1 Phrase Grounding

Phrase grounding aims to locate all of the objects contained in the text query. The relation between query and the generated region is one-to-many. Some of datasets in this formulation are Flickr30K [21] and PhraseCut [22]. An example of the query in this task is ”A man with pierced ears is wearing glasses and an orange hat”. Phrase grounding model will produce 4 bounding box in total, referring ”a man”, ”pierced ears”, ”glasses”, and an orange hat” (See Figure 2.1 (Left)).

### 2.1.2 Specific Visual Grounding or Referring Expression Comprehension

In this type of visual grounding, also known as referring expression comprehension, there is only one specific object in the image referred by the expression query in natural language. As shown in Figure 2.1 (Middle), all necessary information to specify the object is derived from the expression query. One of early works conducted by Yu et al [3], by training a Fast-RCNN [23] detector for VG. They use three dataset to build and evaluate the model, including RefCOCO, RefCOCO+, and RefCOCOg. To locate an object, the model relies on visual attributes described in the query, such as the

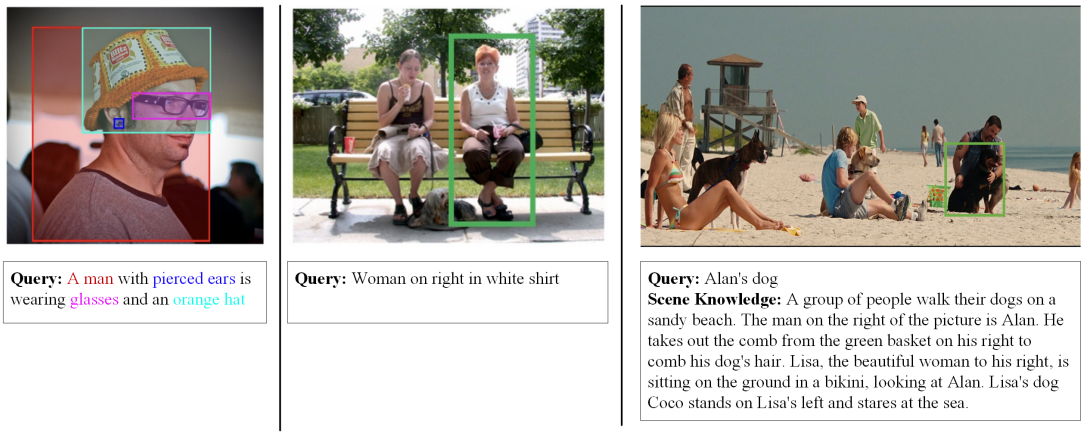



Figure 2.1: **Left**: Phrase grounding with Flickr30k [21], generating multiple bounding box for each phrases in the query. **Middle**: In basic VG on RefCOCO Dataset [3], the model identify exactly one correct bounding box, solely based on physical appearance mentioned in the query. **Right**: In VG with Scene Knowledge [4], deep reasoning ability is needed to understand the given scene knowledge in order to identify the object provided in query.

object type (”girl”, ”orange”, ”dog”), shape (”sharp”, ”rectangular”), position (”back”, ”right”, ”under”), colour (”purple”, ”green”), and pose (”lay”, sit”). Deep reasoning is not required for this task, it is addressed by mapping visual appearances to their corresponding expressions.

### 2.1.3 Visual Grounding with Scene Knowledge

In VG with scene knowledge, the model needs capability to reason with the provided scene knowledge to accurately identify the object referenced in a query. The different with the previous type is shown in Figure 2.1. As described in [4], there are three criteria of the query that differ it from basic VG. First, the phrase is not visually identifiable by appearance, but highly relevant to scene knowledge (knowledge relevance). Second, the phrase is unambiguous (uniqueness). Third, using specific terms such as ”friend” instead of ”person” (diversity). To solve this formulation, [4] proposed two approaches based in the number of stages, illustrated in Figure 2.2.

The first method only consist of one stage, called **Knowledge-embedded Vision-Language Interaction (KeViLI)**. This model is designed from scratch, consists of vision encoder initialized from DETR[24], text encoder initialized from BERT[25], cross attention transformer and self attention transformer. In this method, both query $H_T$

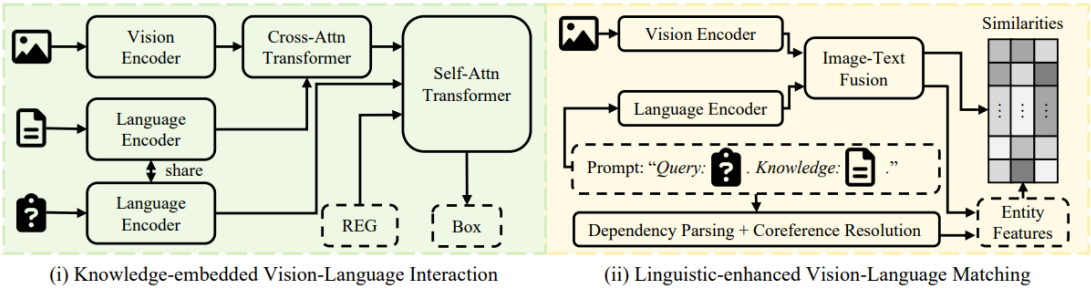



Figure 2.2: KeViLI and LeViLM methods, directly adapted from [4].

and scene knowledge $H_K$ are passed into with a language encoder. Concurrently, image encoder generates the image patch features $H_I$. Then, the image and encoded scene data are embedded using a cross-attention transformer. It consists of cross-attention, self-attention, and feed-forward sub-layers. In the self-attention sublayer, the attention between each image patch is captured. Next, $H_I$ will be the query in the cross-attention sublayer, with both key and value being $H_K$. After that, $H_I$ and $H_T$ together with a learnable token `[REG]`, are fed into transformer to perform image-query interaction. The transformer's output is then passed through MLP layers to generate the coordinate. The loss metrics is computed with following formula:

$$L = L_{\text{smooth\_l1}}(b,\hat{b}) + \mathrm{L}_{\text{giou}}(b,\hat{b}) \tag{2.1}$$

In formula 2.1, $b$ and $\hat{b}$ denotes label and prediction boxes, $L_{\text{smooth\_l1}}(.)$ and $L_{\text{giou\_l1}}(.)$ are the smoothed L1 and GIoU loss.

The second approach, called **Linguistic-enhanced Vision-Language Matching (LeViLM)** involves two phases. The first phase is region proposal stage, finding all objects in the image. The second stage involve the scoring of the proposed regions. In the region proposal stage, a prompt is built as `"Query: T. Knowledge: K."`. This approach uses GLIP[26] as the backbone model, including initialization for language encoder and vision encoder. First, the text prompt and the image are encoded with a language encoder and vision encoder respectively. Then, both representations are passed into image-text fusion through multiple layer. Each layer consist of a self-attention layer to encode the text, a dynamic head layer to encode image, and two cross-attention layers for the fusion. This process generates after fusion image features $Z_I$ and after fusion text features $Z_P$. The detail formulation can be found in [4]. In the region scoring stage, subject is extracted by parsing the structured linguistic information from the query and the scene knowledge, denoted as the head entity $E_h$ . Then, to identify all mentions $E_m$ in the knowledge that correspond to the same entity $E_h$, a connection between the entity

in the query with all of its coreference that mentioned inside the knowledge is developed. Representation of $E_h$ and $E_m$ is taken from $Z_P$, resulting in $Z_E \in R^{(E+1)\times d}$. Finally, they compute the alignment scores between entities in the prompt and the regions with the following formula:

$$Score = Z_I Z_E^T \tag{2.2}$$

The model is trained to minimize the loss function $Loss = L_{xe}(Score, Target)$ where $L_{xe}$ is the cross-entropy loss. Each element in $L$ indicates the matching between region and entity.

## 2.2 One-For-All (OFA)

In recent years, there has been a shift in model development from task specific methods to large scale pre-training. The ”pretrain-finetune” paradigm has demonstrated significant success across various domains. As mentioned in [7], following the advent of BERT[25] in Natural Language Processing (NLP), vision-language research also moved to transformer based model, including UNITER[27], CLIP[28], Flamingo[10], and OFA ”One-for-All”[8]. OFA [8] is a pre-trained model aiming to unify varied modality tasks, including vision-language, vision-only, and language only task. As illustrated in Figure 2.3, it is represented as simple sequence-to-sequence learning framework with common instruction format task representation. OFA is developed to have omnipotent model with following properties:

1. Task Agnostic: has unified task representation supporting multiple task types and agnostic to pretraining or finetuning.

2. Modality Agnostic: Image and text represented as unified format of input and output.

3. Task Comprehensiveness: trained by enough task variety to have generalization ability.

### 2.2.1 Dataset and Input Representation

The pretrain phase of OFA utilizes 20M image-text pairs from dataset of various tasks. List of dataset for each task is shown in Table 2.2. To present multimodality without outputting in task specific schema, texts and images are represented in a common space

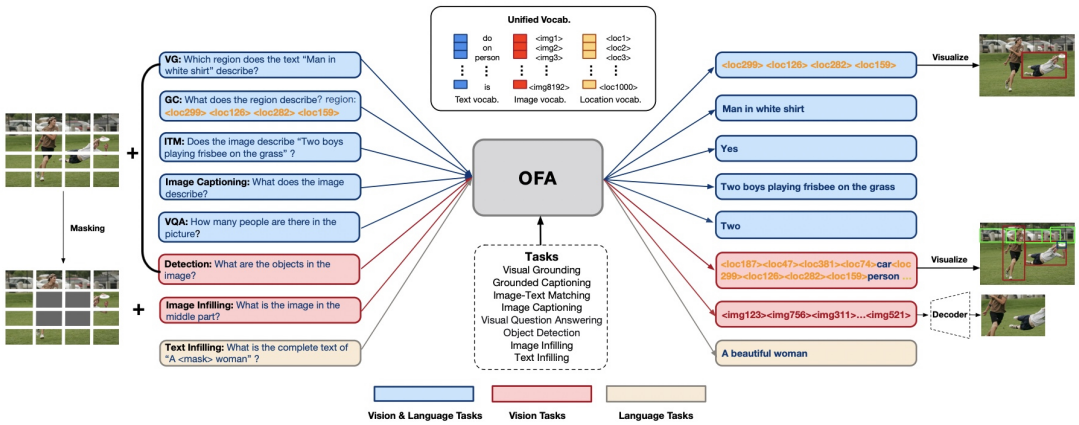



Figure 2.3: High level architecture of OFA, directly adapted from [8].

as tokens in unified vocabulary. For text sequence, byte pair encoding (BPE) is applied. Image is represented as discrete code of smaller area or patches. Location coordinate for bounding box also transformed as location tokens denoting the top left and bottom right coordinate. As tokens, it could be also represented as BPE tokens. All the linguistic and visual tokens are combined in unified vocabulary.

### 2.2.2 Model Architecture and Training

The architecture backbone of OFA is encoder-decoder transformers. The encoder layer comprises self attention and feed-forward network, while decoder layer consists of self attention, FFN, and cross attention as the connection bridge between decoder and encoder output representations. Head scaling is added to self attention, and normalization layer is added after post attention and first layer of FFN in order to stabilize training and accelerate convergence. In term of parameter size, OFA provides five different scale, ranging from $OFA_{Tiny}$ (33M), $OFA_{Medium}$ (93M), $OFA_{Base}$ (182M), $OFA_{Large}$ (472M), $OFA_{Huge}$ (930M).

In order to unify various tasks and modalities, OFA is designed as sequence-to-sequence learning with different modality. Five cross-modality tasks, comprises visual grounding, grounded captioning, image-text matching, image captioning, and visual question answering are used to pretrain OFA. Moreover, OFA is also pretrained with three uni-modal tasks (image infilling, object detection, and text infilling). The unified input and output format for each task is described on Table 2.1. OFA is optimized with cross-entropy loss as presented in Formula 2.3, where $x$ is input, $s$ is instruction and $y$ is

output.

$$\mathcal{L} = -\sum_{i=1}^{|y|} logP_{\theta}(y_i|y < i, x, s) \tag{2.3}$$

Trie-based search, instead of beam search, is utilized as decoding strategy to achieve better quality in generation.

## 2.3 Kosmos-2

Large Languange Models (LLMs), such as GPT [14, 15, 16], LLAMA [18, 19, 20], and PaLM [43] have demonstrated exceptional abilities in handling NLP tasks using zero-shot or few-shot learning. However, they still unintelligible with vision [44]. The integration of LLMs with vision model has given rise to field of Multiodal Large Language Model (MLLMs), an LLM-based model with ability in perceive, reason, and generate outputs using multimodal information. Kosmos-1[45] is one of the earliest work in combining vision data with the large language models (LLMs) into MLLMs. It is trained on massive web-scale multimodal dataset including text corpora, image-caption pairs, and interleaved data of images and texts. Kosmos-1 demonstrated remarkable performance in zero-shot and few-shot setting for various tasks, including language (completion tasks, cloze, commonsense reasoning), vision (image classification), and perception-language (image captioning, webpage question answering, and visual question answering). However, the grounding capability was lack in this version but was added in its subsequent release as Kosmos-2[12].

### 2.3.1 Dataset and Input Representation

To support the new task, GrIT dataset is constructed, containing image-text pairs acquired from COYO-700M[46] and LAION-2B[47] subset. In Kosmos, input is represented in a unified format as a sequence of token. Some special tokens such as <s> and </s> used to indicate start and end of sequence. To denote image, they use <image> and </image>. To support grounding functionality, Kosmos-2 represents location as token. Initially, the image's height and width are divided into $P$ segments, creating $PxP$ discrete region. The bounding box is defined by the top-left and bottom-right points. As a result, the grounded input representation in Kosmos-2 is structured as  TextSpan <box><loc$_{upperleft}$><loc$_{bottomright}$></box>. To indicate beginning and end of phrase,  and  are used, while <box> signifies the bounding box associated with

| **Type** | **Pretraining Task** | **Input** | **Output** |
|---|---|---|---|
| cross-modal | visual grounding | image and instruction ”Which region does the text xt” describe | bounding box in format <x1,y1,x2,y2> |
| cross-modal | grounded captioning | image and instruction ”What does the region describe? region: <x1,y1,x2,y2>” | region caption |
| cross-modal | image-text matching | image and instruction ”Does the image describe xt?” | ”Yes” or ”No” |
| cross-modal | image captioning | image and instruction ”What does the image describe?” | image caption |
| cross-modal | visual question answering | image and question | answer |
| uni-modal | image infilling | image with masked middle part and instruction ”What is the image in the middle part?” | sparse code for the masked middle part |
| uni-modal | object detection | image and instruction ”What are the objects in the image?” | sequence of object position and label |
| uni-modal | text infilling | text with masked part | masked part text |

Table 2.1: Tasks and its input and output to pretrain OFA, adapted from [8].

| **Type** | **Pretraining Task** | **Dataset Source** |
|---|---|---|
| Vision & Language | Image Captioning, Image-Text Matching | CC12M [29], CC3M[30], SBU[31], COCO[32], VG-Cap[33] |
| | Visual Question Answering | VQAv2[34], VG-QA[33], GQA[35] |
| | Visual Grounding, Grounded Captioning | RefCOCO[3], RefCOCO+[3], RefCOCOg[36], VG-Cap[37] |
| Vision | Detection | OpenImages[38], Object365[39], VG[33], COCO[32] |
| | Image Infilling | OpenImages[38], YFCC100M[40], ImageNet-21K[41] |
| Language | Masked Language Modeling | Pile (Filtered)[42] |

Table 2.2: Dataset sources for OFA pretraining, adapted from [8].

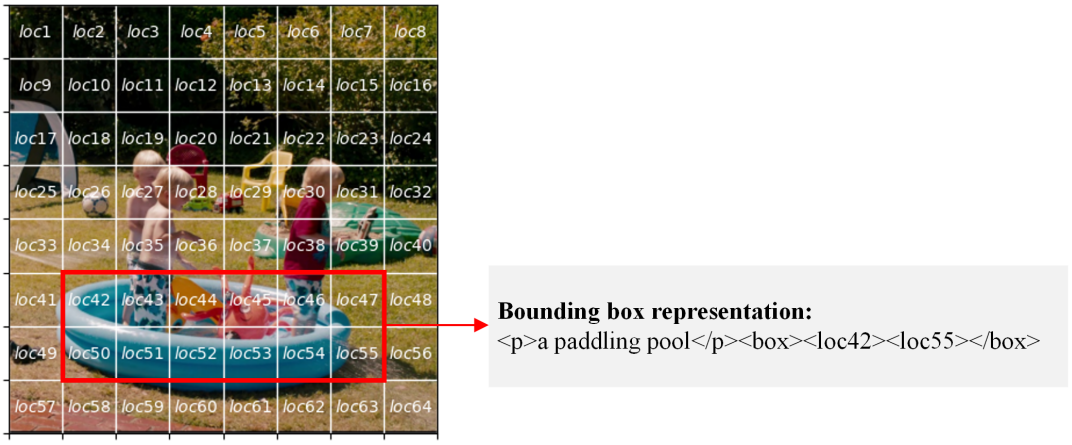



Figure 2.4: Illustration of how to transform the bounding box as token input representation. The upper left coordinate for the paddling pool is $loc_{42}$ and the bottom right coordinate is $loc_{55}$. This example uses 8x8 grid for simplification. The actual implementation divides the image into 32x32 location, resulting in 1024 different location tokens.

its text span. Table 2.3 explained some format representation used to train Kosmos-2. Figure 2.4 illustrates the input representation of phrase and its bounding box.

### 2.3.2 Model Architecture and Training

Kosmos-2 follows a training approach closely aligned to Kosmos-1, which was constructed using the MetaLM [48] framework. This model is a transformer-based language model that integrates the vision modules in its architecture. Kosmos-2 utilizes TorchScale [49] as its base library, acts as the foundation of Transformer variant Magneto [50]. For position encoder, xPos [51] is used. Figure 2.5 illustrates the high level architecture of Kosmos-2. First, a unified representation is constructed from both text and image inputs before being processed into the Transformer-based decoder. The model generates sequences in an auto-regressive manner, where predictions of subsequent tokens are influenced by the preceding timesteps tokens. Causal mask is applied to prevent future information leakage. Finally, a softmax layer is employed to choose tokens from the predefined vocabulary. This architecture showcases flexibility in handling vision-language data, as long as the inputs can be encoded as vectors.

As a next prediction token task, the training process aims to maximize the log-likelihood of tokens present in the dataset. The vision encoder module consists of 24 layers of 2024 hidden size, and Feed Forward Networks (FFNs) of size 4096. This

| Model | Prompt Format |
|---|---|
| Text | <s>{Text}</s> |
| Image-Caption | <s> <image> {ImageEmbedding} </image> {Caption} </s> |
| Interleaved Image-Text | <s> <image> {ImageEmbedding} </image> {Caption1} {ImageEmbedding2} </image> {Caption2} </s> |
| Grounding | <s> <image> {ImageEmbedding} </image> <grounding>  It  <box> <$loc_{topleft}$> <$loc_{bottomright}$> </box> is {expression} </s> |

Table 2.3: Data input representation in Kosmos-2

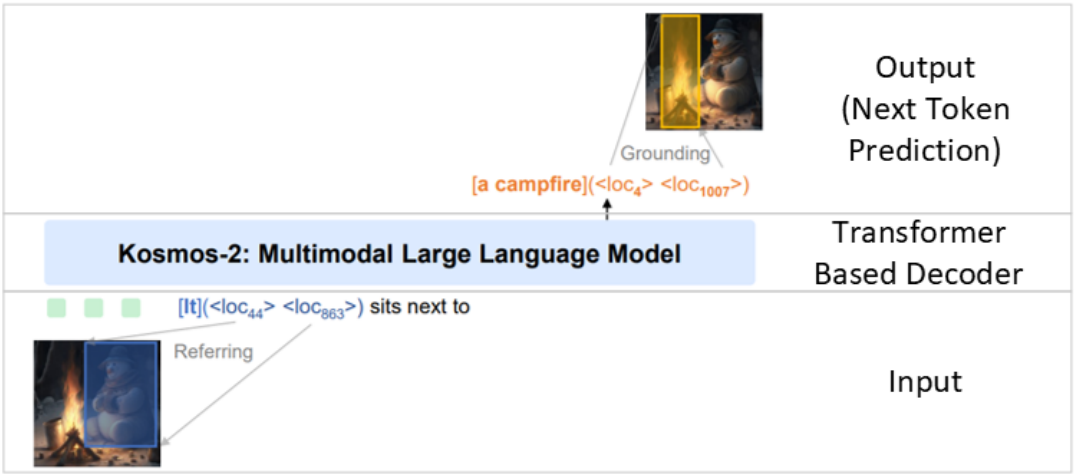



Figure 2.5: High level architecture of Kosmos-2, directly adapted from [12] with additional note.

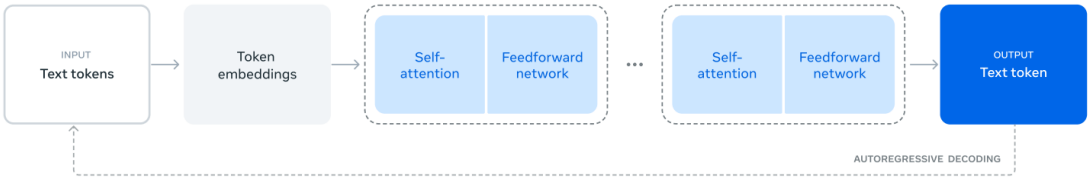



Figure 2.6: Overall architecture and training of Llama3, directly adapted from [20].

framework employ 24 layer Magneto transformer as MLLM part, featuring 32 attention heads, an intermediate FFN size of 8192, and 2048 hidden dimensions. Overall, the model contains 1.6B parameters, with weight initialization from Kosmos-1. To enhance the alignment with human instructions, the model integrates two instruction datasets: a vision-language instruction dataset LLaVA-Instruct [9] and the language-only instruction dataset FLANv2 [52]. Additionally, specific grounded instruction is incorporated to refine model capability. It consists of bounding box pairs and expressions from GrIT.

## 2.4 Llama 3

Llama 3 [20] is a large language foundation model that supports a broad set of capabilities, including solving complex reasoning problem, reading comprehension, code generation, multilinguality, and many more. Figure 2.6 presented the overall architecture and training of Llama 3. As mentioned in [20], there are two main stages in Llama3 development:

- **Pre-training.** In this phase, the model is trained with 15.6T tokens of large multilingual text corpus using next token prediction task. During pre-training, the model learns language structures and acquires world knowledge. The dataset is curated from various web data with de-deplication, cleansing, and the removal of personal identifiable information (PII) and adult content. Architecturely, Llama 3 closedly follows Lllama [18] and Lllama2 [19], which are based on Transformer [53].

- **Post-training.** In order to make LLM behaved well according to the instructions, Lllama3 is further applied with six round of post-training and enhanced with human feedback on top of the pre-trained checkpoint. These rounds include supervised fine-tuning and direct preference optimization [54]. The post-training data consists of specially targeted synthetic data, human annotations with rejection-sampled responses, and human-curated datasets.

Llama 3 is available in various sizes, with parameter number ranging from 8 billion, 70 billion, to 405 billion. It also comes in two version (”Base” and ”Instruct”), depending on whether it has been fine-tuned with instructions.

# Chapter 3

# Dataset

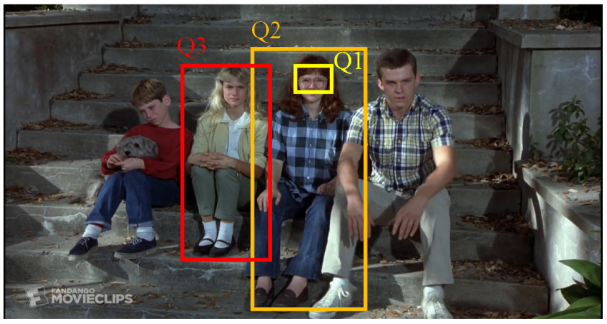
## 3.1 Dataset Description

To assess the capabilities of visual grounding combined with reasoning tasks, we utilize the SK-VG dataset. It contains 4000 pairs of images, scene knowledge, queries, and bounding boxes. This dataset consists of movies scenes from Visual Commonsense Reasoning dataset [55] with some filtering criteria. As explained in [4], the scene must have human as the main body of the story, have objects to describe the details of scene, and the scene location background (e.g. park, classroom, beach). For the query, as the dataset is built to evaluate VG with reasoning, it must be differ from traditional VG query. There are 3 criteria for the expression:

1. Knowledge relevance: must be indirectly derived from the story rather than visually distinguishable.
2. Uniqueness: should only refer to exactly one object or region in the image.
3. Diversity: general lexicon such as ”person” is replaced with more specific terms (”colleague”, ”girl”, ”teacher”).

## 3.2 Dataset Splits

The dataset is splitted into train (2304 samples), validation (9180 samples) and test (6598 samples). The test set is further categorized into three difficulty levels based on the visual distinguishability of the expressions: easy (3028 samples, 45.89%), medium (1828 samples, 27.71%), and hard (1742 samples, 26.30%). In the easy example, the




**Query 1 (Easy):**
The glasses worn by Lisa

**Query 2 (Medium):**
The sister named Lisa

**Query 3 (Hard):**
The person who is too tired to speak

**Scene Knowledge:**
The man on the far right of the image is Mark. He takes his family out to climb the mountain. His sister Lisa is sitting on Mark's right with glasses. Ann has blonde hair and sits on Lisa's right, too tired to speak. Alan, dressed in red, sits on Ann's left and holds his pet dog Coco.

Figure 3.1: The examples of three different levels of query difficulty. In this particular example, the easy query does not need much reasoning since there is only one glasses in the image. For the medium level, we need to look into the story and look for visual description of Lisa. In the hardest level, the subject is not directly mentioned. It requires more reasoning to know who is ”too tired to speak” from the knowledge, and then look for visual (blonde hair) and spatial description (Lisa’s right) to detect the right object.

query contains obvious object or visual description. For medium, it contains weak visual clues. Last, the scene knowledge is fully needed to derive the answer for the hard difficulty.

The comparison of different difficulty is presented in Figure 3.1. In this particular example, query 1 is classified as easy since there is only glasses in the image, the proper noun ”Lisa” is not necessary to detect correct bounding box. Expression ”The sister named Lisa” in query 2 is considered medium because the noun ”sister” alone is not enough to identify the correct region, as there are two women in the image. In the hard difficulty example, multiple steps reasoning is needed to accurately choose the referred person. First, we need to know ”who is too tired to speak”, which is Ann. Second, we look for Ann’s description to pick the right person, either by visual (”Ann has blonde hair”), or position(”sits on Lisa’s right”).

# Chapter 4

# Methodology

In this chapter, we explained the preprocessing and evaluation method in comparing Kosmos-2 and OFA for visual grounding with reasoning requrement. The explanation for each evaluation settings is provided in Section 4.3.

## 4.1 Image Preprocessing

Kosmos-2 encoder is built on CLIP image processor which resize the image into the size of 224 x 224. For OFA, each image is resize to 512 x 512.

## 4.2 Evaluation Metric

To evaluate the system, we use the Intersection over Union (IoU), which calculates the fraction between the region of intersection and the region of union between the predicted bounding box and the labeled bounding box. It is mathematically defined in Equation 4.1.

$$IoU(Prediction, Label) = \frac{Prediction \cap Label}{Prediction \cup Label} \tag{4.1}$$

To compare the accuracy of each model configuration, we choose IoU threshold of 0.5, which means the prediction is considered correct if the IoU greater or equal to 0.5 and considered incorrect otherwise.

## 4.3 Evaluation Settings

In this project, we compare the performance of VLMs under two different settings: zero-shot and fine-tuned. The variations of input format in each setting are explained

| Model | Prompt Format |
|---|---|
| Kosmos-2 | <grounding><phrase>{query}</phrase> |
| Kosmos-2 | <grounding>Knowledge:{scene knowledge} <phrase>{query}</phrase> |
| Kosmos-2 | <grounding><phrase>{revised query}</phrase> |
| OFA | Which region does the text "{query}" describe? |
| OFA | Which region does the text "{knowledge} {query}" describe? |
| OFA | Which region does the text "{revised query}" describe? |

Table 4.1: Zero shot prompt format.

in each section below. We do not try few-shot approach because there is no available implementation code for Kosmos 2. For OFA, the author in [8] said that few-shot in-context learning is not possible due to its model and training design.

### 4.3.1 Zero Shot

With zero-shot learning, we intend to investigate model's ability to perform visual grounding without ever being exposed to the SKVG dataset, relying solely on the data it was trained on during the pretraining phase. We compare the zero-shot capability of Kosmos-2 and OFA in three different condition based on the text input to the model.

#### 4.3.1.1 Zero Shot with Query Only

Merely providing the query in VG with SKVG dataset is very difficut, even impossible for human to identify the correct bounding box in certain cases. As illustrated in Figure 1.1, even human will not be able to know which person is Alan without comprehending more information from the scene knowledge. Nonetheless, we conduct this experiment in order to examine the significance of scene knowledge in enabling the model to accurately determining the bounding box.

#### 4.3.1.2 Zero Shot with Query and Knowledge

In order to understand the impact of knowledge on the model's ability to accurately detect the bounding box, we conduct a comparative experiment with and without the inclusion of additional knowledge inside the prompt. Since there is no designated prompt

template in Kosmos-2 for visual grounding with scene knowledge, we modify the existing example prompts as shown in Table 4.1. For OFA, we simply concat the knowledge and the query with format "`Knowledge: {knowledge} Query:{query}`".

#### 4.3.1.3 Zero Shot with Modified Query

Since Kosmos-2 and OFA are not inherently designed to handle VG task with reasoning capability, we modify the query in order to make it distinguishable by visual description or relative position. This modification would replace all pronouns based on visual cues provided in the scene knowledge. As depicted in Figure 4.1, ”Alan” would be replaced by its descriptive attributes, such as appearance (”the man”), action (”take out comb”), or relative spatial location (”on the right picture”). To extract such information and transform the expression, we do prompting with Large Language Model, particularly LLAMA3-8B [18]. After that, we apply post-processing to eliminate some unnecessary parts of the LLM's response, retaining only the main intended answer, which is the modified expression. Finally, we pass the revised query to the VLM and retrieve the bounding box.

In building the prompt, we also try few shot learning by giving model some examples at inference time without any gradient update, as shown in Figure 4.2. As suggested in [16], few-shot setting in GPT-3 nearly matching some of the state-of-the-art fine tuned systems. Giving few reduce ambiguity for model and help to understand specific answer format. Llama also has shown promising few-shot results in many language tasks [18, 19, 20]. To investigate the effect of including examples in Llama 3 on modified query quality, we compare the grounding performance of both modified queries (with or without examples in the Llama prompt) in Section 5.1.4.

### 4.3.2 Fine-Tuned Model

The pretrain-finetune paradigm is a widely used approach in machine learning, where a large pretrained model is further trained on a task specific dataset. In contrast to previous approaches where the model has not been exposed to the SKVG dataset, this approach involves further training of the pretrained model with SKVG training data. To analyze the impact of incorporating a reasoning pipeline with LLM, we fine-tune the OFA pretrained model under two different conditions. First, we pass the original scene knowledge and the query with format "`Knowledge: scene knowledge. Query: query`" to investigate the OFA model's ability to adapt to this new prompt format.

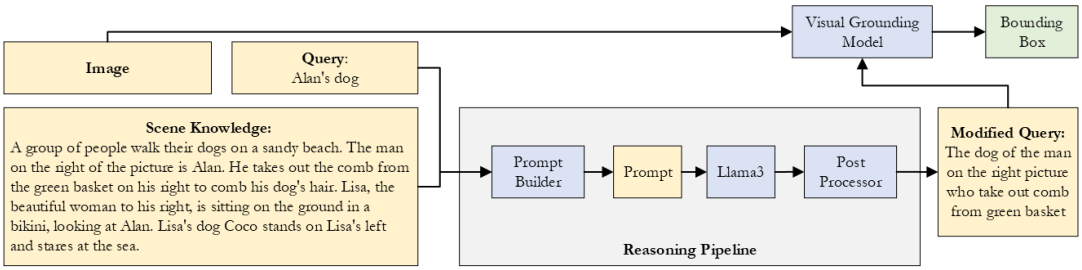



Figure 4.1: We propose to add a reasoning pipeline to let the model comprehend subject identity to modify the query before VG phase. In this particular example, the model need to understand who ”Alan” is first then replace it with the descriptive attributes, such as appearance (”the man”), action (”take out comb”), or relative spatial location (”on the right picture”).

```
I have a task for you to transform the expression that will help a reference expression comprehension model easily
detect the bounding box.
Use visual descriptions or spatial relationships. Prioritize visual description over spatial relation.
Do not generating proper noun, such as person name in the transformed expression.
Let's think step by step.

Here some examples to do it:
1. Knowledge Context: The business man Peter is in the far left of image. He wears glasses and carries a briefcase in
his hand. When he wants to go home after work, a man with white hair goes ahead of Tom and stops him. The man is David
and wears a tie. He loses his wallet and asks Peter if he has seen his wallet.
Expression to Transform: The glasses worn by Peter
Transformed expression: The glasses worn by business man who wears glasses and carries a briefcase.
2. Knowledge Context: Mike in a pink shirt is sitting at the round table and talking to his girlfriend lying on the
sofa. Bob in a white shirt comes to the table and sits in the chair. And he asks Mike why he hasn't returned his
money. And then they quarrel about it.
Expression to Transform: The boy arguing with Mike
Transformed expression: The boy in white shirt

Now, based on the following example, please transform this expression:
Knowledge Context: {knowledge}
Expression to Transform: {query}
Transformed expression:
Answer with the transformed query only.
```

Figure 4.2: Prompt format for LLAMA3 as part of reasoning pipeline to revise the query expression.

Second, we utilize a modified query format as explained in Section 4.3.1.3.

#### 4.3.2.1 Fine-Tuning Configuration

Following the implementation setting in [8], we fine tune the model for 12 epochs with learning rate of $3x10^{-}5$, warm up ratio of 0.06, label smoothing of 0.1, and drop out rate of 0.1. More comprehensive list of hyperparameters can be found in the Appendix A.

# Chapter 5

# Results and Discussion

## 5.1 Zero Shot

### 5.1.1 Query Only

As indicated in Table 5.1, Kosmos-2, $\text{OFA}_{\text{Large}}$ (ID: ZS9), and $\text{OFA}_{\text{Huge}}$ (ID: ZS12) exhibit superior performance compared to LeViLM (ID: ZS1) in zero-shot with query only setting. For $\text{OFA}_{\text{Base}}$, it achieved similar overall accuracy (29.17) compared to LeViLM (29.75). Interestingly, when considering difficulty levels, $\text{OFA}_{\text{Base}}$ performed slightly better in medium cases and achieved approximately 75% higher accuracy in hard difficulty.

In zero-shot setting, model size can be a significant factor affecting performance. As discussed in [4], LeViLM is built following GLIP [26] as its backbone. Comparing the model sizes, as shown in Figure 5.1, Kosmos-2, $\text{OFA}_{\text{Large}}$, and $\text{OFA}_{\text{Huge}}$ are larger than GLIP [1].

With more parameters, generally model possess a higher capacity to learn complex pattern from data. The relation between model capacity and its accuracy is clearly reflected in three version of OFA with different model size. Surprisingly, despite having fewer parameters, $\text{OFA}_{\text{Huge}}$ (930M parameters) outperforms Kosmos-2 (1660M parameters).

Model size is not the only factor affecting the performance, the dataset and type of task during initial pretraining also plays crucial roles. GLIP is pretrained with phrase grounding tasks, aims to detect multiple bounding boxes (not single) for each detected

[1] Since [4] does not specify which version of GLIP is used in LeViLM, we assume it is GLIP-Tiny with 232M parameters.

| VG Model | ID | Text | Overall Acc | Difficulty Level | | |
|---|---|---|---|---|---|---|
| | | | | $\mathbf{Acc_{Easy}}$ | $\mathbf{Acc_{Medium}}$ | $\mathbf{Acc_{Hard}}$ |
| **Zero Shot** | | | | | | |
| LeViLM | ZS1 | Q | 29.75 | 49.97 | 18.23 | 6.71 |
| | ZS2 | Q+K | 7.55 | 13.08 | 4.38 | 1.26 |
| Kosmos-2 | ZS3 | Q | 38.63 | 49.01 | 32.66 | 26.87 |
| | ZS4 | Q+K | 39.89 | 47.46 | 34.41 | 32.50 |
| | ZS5 | MQ | 40.03 | 48.08 | 34.85 | 31.46 |
| $OFA_{Base}$ | ZS6 | Q | 29.17 | 39.43 | 22.92 | 17.91 |
| | ZS7 | Q+K | 16.06 | 13.97 | 16.03 | 19.75 |
| | ZS8 | MQ | 33.28 | 41.11 | 28.12 | 25.09 |
| $OFA_{Large}$ | ZS9 | Q | 42.47 | 58.09 | 33.21 | 25.03 |
| | ZS10 | Q+K | 18.72 | 17.24 | 18.05 | 21.99 |
| | ZS11 | MQ | 53.30 | 63.11 | 46.78 | 43.11 |
| $OFA_{Huge}$ | ZS12 | Q | 48.14 | 64.60 | 39.28 | 28.82 |
| | ZS13 | Q+K | 20.93 | 17.01 | 21.39 | 27.27 |
| | ZS14 | MQ | **57.85** | **66.08** | **51.04** | **50.69** |
| **Fine Tuned** | | | | | | |
| LeViLM | FT1 | Q | 57.18 | 80.35 | 46.80 | 27.83 |
| | FT2 | Q+K | 70.70 | 84.51 | 63.16 | 54.62 |
| | FT3 | Q+K+S | **72.57** | **83.72** | **65.52** | 59.95 |
| $OFA_{Base}$ | FT4 | Q+K | 53.15 | 69.35 | 44.26 | 34.33 |
| | FT5 | MQ | 59.56 | 66.55 | 51.86 | 55.51 |
| $OFA_{Large}$ | FT6 | MQ | 69.40 | 77.77 | 63.51 | 61.02 |
| $OFA_{Huge}$ | FT7 | MQ | 70.11 | 78.37 | 63.57 | **62.63** |

Table 5.1: Performance of Kosmos-2 and OFA towards SKVG Dataset in zero-shot and fine tuned setting. In zero-shot, there are three different text input. Q represents query only, Q+K represents concatenation of query and knowledge, and MQ represents modified query. In fine tuned setting with LeViLM, S is linguistic structure. The modified query is generated by Llama 3, utilizing its reasoning capability given the original query and the scene knowledge.

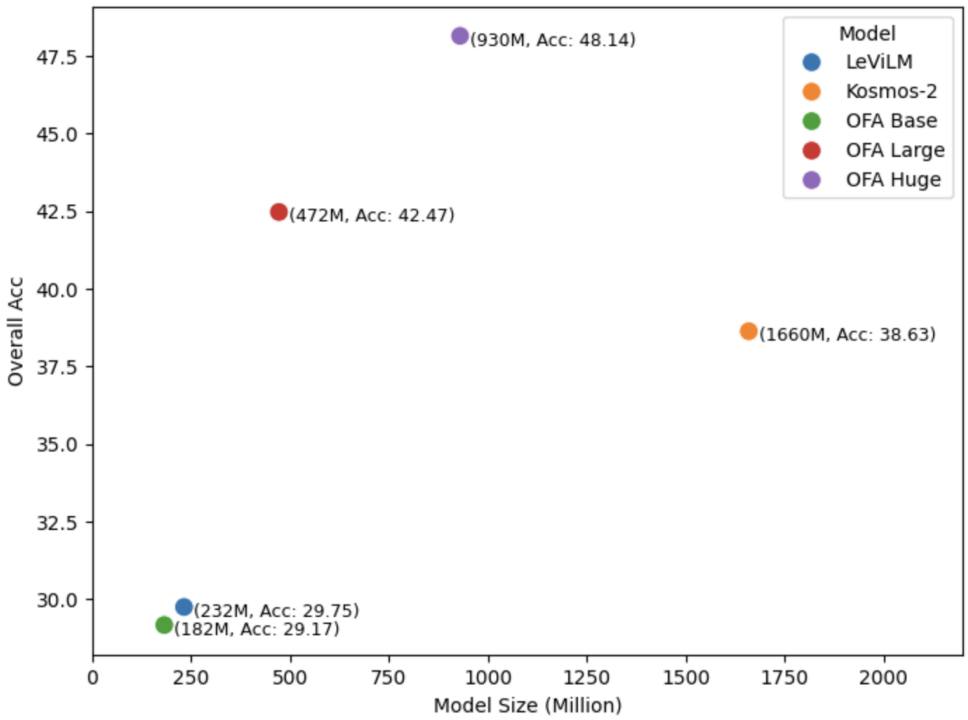



Figure 5.1: Model Size Comparison and Accuracy.

object mentioned in text queries. It is differs from visual grounding (also mentioned as referring expression comprehension) task where the goal is identify exactly one bounding box referred by the expression. The object phrases in GLIP are relatively shorter and simple than SK-VG dataset, such as ”blow dryer”, ”protective goggles”, or ”beautiful carribean sea turqoise”. In contrast, Kosmos-2 is trained with one-to-one pair of descriptive phrases (”a dog in a field of flowers”) and bounding boxes. OFA is also pretrained with visual grounding task utilizing RefCOCO, RefCOCO+, and RefCOCOg. Those datasets feature varied and linguistically rich examples, including sentences with visually descriptive phrases (”white shirt guys”), spatial position (”building on right behind guys”), or actions (”the little kids holding a racket”). This exposure to diverse and complex expressions likely contribute to OFA’s superiority compared to other models.

### 5.1.2 Adding Knowledge by Simple Concatenation

Similar with LeViLM [4], OFA struggles to detect bounding box accurately when the knowledge is added by simple concatenation (ID: ZS7, ZS10, ZS13). The performance in this scenario is worse than using the query alone. This does not imply that knowledge harm the model, however the addition of knowledge results in a much longer text prompt

compared to the text representation in OFA's pretraining dataset. The average number of words in the concatenated query is 65.62. In contrast, OFA VG pretraining task uses RefCOCO[3], RefCOCO+[3], RefCOCOg[36], and VG-Cap [37] with only 3.5, 3.5, 8.4, and 5 words per query expression respectively, which are significantly shorter than text in the Q+K setting. Interestingly, this phenomenon is not happen in Kosmos-2.

### 5.1.3 With Modified Query Generated by LLM Reasoning

As shown in Table 5.1, Kosmos-2 and OFA achieved superior performance compared to LeViLM in zero-shot setting through query modification. Query rewriting slightly increased Kosmos-2 (ID: ZS5) accuracy (3.5%) and significantly boosted the accuracy for OFA (ID: ZS8, ZS11, ZS14). The overall accuracy improvements were 14.09% for $\text{OFA}_{\text{Base}}$, 25.5% for $\text{OFA}_{\text{Large}}$, and 20.17% for $\text{OFA}_{\text{Huge}}$.

As shown in some examples depicted in Table 5.2, the reasoning capability of Llama 3 successfully revise the queries to exclude any proper nouns, making objects easier to identify. It is replaced with visual descriptions ("in black suit" in example 1) or spatial relationships ("Mark" is rewritten as someone who "standing on the left of the image"). However, we argue that some phrases in the modified query is unnecessary and could be erased in order to get shorter prompts. For example, in example number 3, the description "the person with yellow hair" is specific enough to describe the referred person without the needed of phrase "being protected by the person with his arm around him". We also found that Llama 3 failed to correctly modify some queries. For instance, in example 4 of Table 5.2, the pronoun "Mark" should be replaced as "The person who sits on the chair opposite someone kneeling". Table 5.3 presents the percentage of revised queries that still contain proper nouns. As expected, Llama 3 struggles with queries of high difficulty, with proper nouns remaining in 19.35% of the hard samples, followed by 14.62% in medium difficulty, and 8.62% in easy category.

### 5.1.4 Effect of Providing Examples in Llama 3 Prompting

In this section, we discuss in-context learning by providing examples within prompt for Llama 3. Table 5.4 presents the accuracy comparison of modified query generated by Llama 3 with and without examples. In both scenarios, the Llama 3 was instructed to transform the query as shown in Figure 4.2, with additional guidance such as "using visual description", "avoiding proper noun", and "thinking step by step". The key difference between the two setups is the presence of example. As stated in table 5.4,

| ID | | Knowledge, Query, and Modified Query |
|---|---|---|
| 1 | K | The housekeeper Danny, who is a man in a black suit, comes out of the house. The woman in front of him, Sunny, is walking towards her husband Leon, who is wearing a gray suit. Leon has just come home from work when he sees his wife walking towards him. He feels very happy. |
| | Q | The man who is a housekeeper |
| | MQ | The man in black suit |
| | Q | The person who has just come home |
| | MQ | The person in gray suit who feels very happy. |
| 2 | K | On the training ground, the officer Mark, standing on the left of the image with a gun in his hand, arranges for Kevin to check why the soldier Alan in front of Mark is lying on the ground. Kevin kneels next to Alan on one knee, checking Alan's body. Jack, the soldier standing on Mark's right, is Alan's good friend and is very worried about Alan's safety. |
| | Q | The officer named Mark |
| | MQ | The officer standing on the left of the image with a gun in his hand |
| | Q | The helmet worn by Jack |
| | MQ | The helmet worn by soldier standing on the right of officer with a gun |
| 3 | K | Alan, a little boy with yellow hair on the middle of the image, is isolated. Eric, wearing a gray vest behind Alan, keeps mocking him behind him. Alan's good friend Owen puts his arm around Alan on his left, trying to protect him. But Frank, who is wearing a black coat to Owen's right, wants to stop them. |
| | Q | The person being protected by Owen |
| | MQ | The person with yellow hair being protected by the person with his arm around him. |
| 4 | K | The piano is on the left side of the image. Mark kneels on the right side of the piano and is taking off his hat. He had just performed to Nick, who is sitting across from him, and is greeting Nick at this time. Nick sits on the chair opposite Mark and has watched the whole Mark's performance. He is satisfied with Mark's performance. |
| | Q | The person who is satisfied with Mark's performance |
| | MQ | The person who sits on the chair opposite Mark |

Table 5.2: Some examples of modified query generated from Llama 3. K represents knowledge, Q represents original query, and MQ represents modified query.

| Difficulty | Number of Sample | Number of Sample Containing Pronoun | |
|---|---|---|---|
| | | 2 Examples Provided | No Example Provided |
| Easy | 3028 | 261 (8.62%) | 776 (25.63%) |
| Medium | 1828 | 249 (13.62%) | 467 (25.55%) |
| Hard | 1742 | 337 (19.35%) | 524 (30.08%) |

Table 5.3: Percentage of modified query that still containing proper noun. We compare effect of providing examples in the Llama 3 prompt.

| VG Model | Text | Overall Acc | Difficulty Level | | |
|---|---|---|---|---|---|
| | | | $\text{Acc}_{\text{Easy}}$ | $\text{Acc}_{\text{Medium}}$ | $\text{Acc}_{\text{Hard}}$ |
| $\text{OFA}_{\text{Huge}}$ | Modified query without providing example in prompt | 47.91 | 48.25 | 46.55 | 48.74 |
| $\text{OFA}_{\text{Huge}}$ | Modified query with 2 examples in prompt | 57.85 | 66.08 | 51.04 | 50.69 |

Table 5.4: Impact of including examples on grounding accuracy.

the overall accuracy significantly drops by around 10 points. The most notable decline occurs in the easy category. One indicator of inaccurate query rewriting is the presence of proper noun. Table 5.3 reveals that the number of expression that still containing proper noun in the easy category increase substantially when no examples are provided. It jumped from 8.62% to 25.62%, the highest increase compared to the medium and hard category. We can conclude that providing examples significantly improve Llama 3's responses. Examples clarify what exactly is being asked, reduce ambiguity, and help model to understand intended format or specific requirement. This finding aligns with findings in [18] and [16].

## 5.2 Fine Tuning

As explained in Section 4.3.2, we compared two different approaches in our fine-tuning efforts. The first approach involved providing the model with both the query and the knowledge to investigate whether the model has capability to learn the relationship between the query and the knowledge, as well as the reasoning ability to solve the

grounding problem. In this approach, the VG model is responsible for performing the reasoning task. In the second approach, the reasoning task is handled by an LLM, allowing the OFA model to focus solely on the grounding task tuning.

### 5.2.1 Fine Tuning by Providing Query and Knowledge

As shown in Table 5.1, compared to the zero-shot approach with the modified query of $OFA_{Base}$ (ID: ZS8), the fine-tuned model with concatenation (Q+K) (ID: FT4) achieved an improvement in accuracy by 19.87 points. This aligns with the downstream tasks results presented in OFA paper [8], which demonstrate that the OFA pretrained model have general understanding of texts and images, and can effectively adapt to new instructions with fine-tuning. In our case, the pretrained model is able to learn new format of prompt and recognizing two keywords (”Knowledge” and ”Query”). It proved that OFA could have a reasoning ability to understand the scene knowledge story and choosing the correct bounding box.

### 5.2.2 Fine Tuning Using Modified Query

Although the fine-tuned OFA has shown its reasoning ability to understand the knowledge as mentioned in Section 5.2.1, it does not match the LLM reasoning ability, in this case Llama 3. As presented in Table 5.1, fine-tuning with a modified query using $OFA_{Base}$ (ID: FT5) outperforms the Q+K fine tuning technique (ID: FT4) by 6.41 points. Llama3-8B-Instruct, with 8 billion parameters, has been pretrained on over 15T tokens, and finetuned with human feedback, indeed has superior capability in reasoning.

In comparison with the fine-tuned version of LeViLM, our fine-tuned $OFA_{Huge}$ (ID: FT7) achieves competitive result against the LeViLM Q+K (ID: FT2). While FT7 accuracy lags behind FT2 by around 6 points in easy category, it exhibits similar result in medium category and shows significant superiority in hard category by around 8 points. It is also important to note that LeViLM is a specialized model, specifically designed for visual grounding, whereas OFA is built as general purpose model that we fine-tuned for visual grounding task.

Overall, the fine-tuned version of LeViLM involving linguistic structure (ID: FT3) achieved the best accuracy, including the easy and medium category. Notably, our fine-tuned $OFA_{Huge}$ (ID: FT7) holds its superiority in hard category. We argue that the absence involvement of linguistic structure may contribute to its misfocusing the correct bounding box in certain cases, showing its struggle to find the main noun of the

expression. This issue is further discussed in Section 5.3, along with Figure 5.6 (row 2-3).

## 5.3 Qualitative Analysis

To further investigate the effects of query modification query and fine-tuning, we perform a qualitative analysis on SK-VG dataset. We focus on $\text{OFA}_{\text{Large}}$ for the comparison with some results presented in Figure 5.2. In first row, VG model with all three settings correctly detects the bounding box. Despite the presence of proper noun ”Carl” in the original query, our model accurately predicts the bounding box. This accuracy is likely because the expression contain word the ”daughter” and there is only one woman in the image. In the second case (row 2), the expression refer to a hat worn by waiter Leon. However, since multiple hats are present, the model needs to pick one. The zero shot attempt with the original query fails to identify the correct hat. With the help from Llama 3 reasoning, the query is modified to make the referred hat is less ambiguous. The modification replaces ”Leon” with ”the waiter in the middle of image”, allowing model to accurately detect the hat based on its spatial position. In row 3, both zero-shot setting fail, but fine-tuned model succeeds. In the fine-tuned version, the model has been trained with similar types of image (movie scenes) along with the expressions. Finally, in the fourth row, all model failed. The Llama 3 reasoning also generates a wrong rewriting by producing another proper noun (”Abby”).

Query modification using Llama 3 improve the model performance significantly, particularly for categorically hard samples. We selected several hard samples and present the grounding results in Figure 5.3. We notice that the majority of improvement is due to expression transformation of ambiguous or abstract phrases into more concrete visual description, such as what the person is wearing or their physical attributes. As shown in Figure 5.3, the abstract phrase ”has just come home” is rewrite as ”the person in gray”, which make it trivial for grounding model to comprehend. Similarly, the phrase ”too tired to speak”, which is hard to imagined visually, is transformed into ”person with blonde hair”, making it easier for model to interpret.

On the other hand, we also find that some samples from easy category which are correctly predicted with original query, become incorrect after query modification. The issue arises from inaccuracy in reasoning pipeline. As presented in Figure 5.4, query rewriting can change the focus of the expression. This shift is likely to occur if the head word of the expression is an object, rather than a person. We observe that there is a

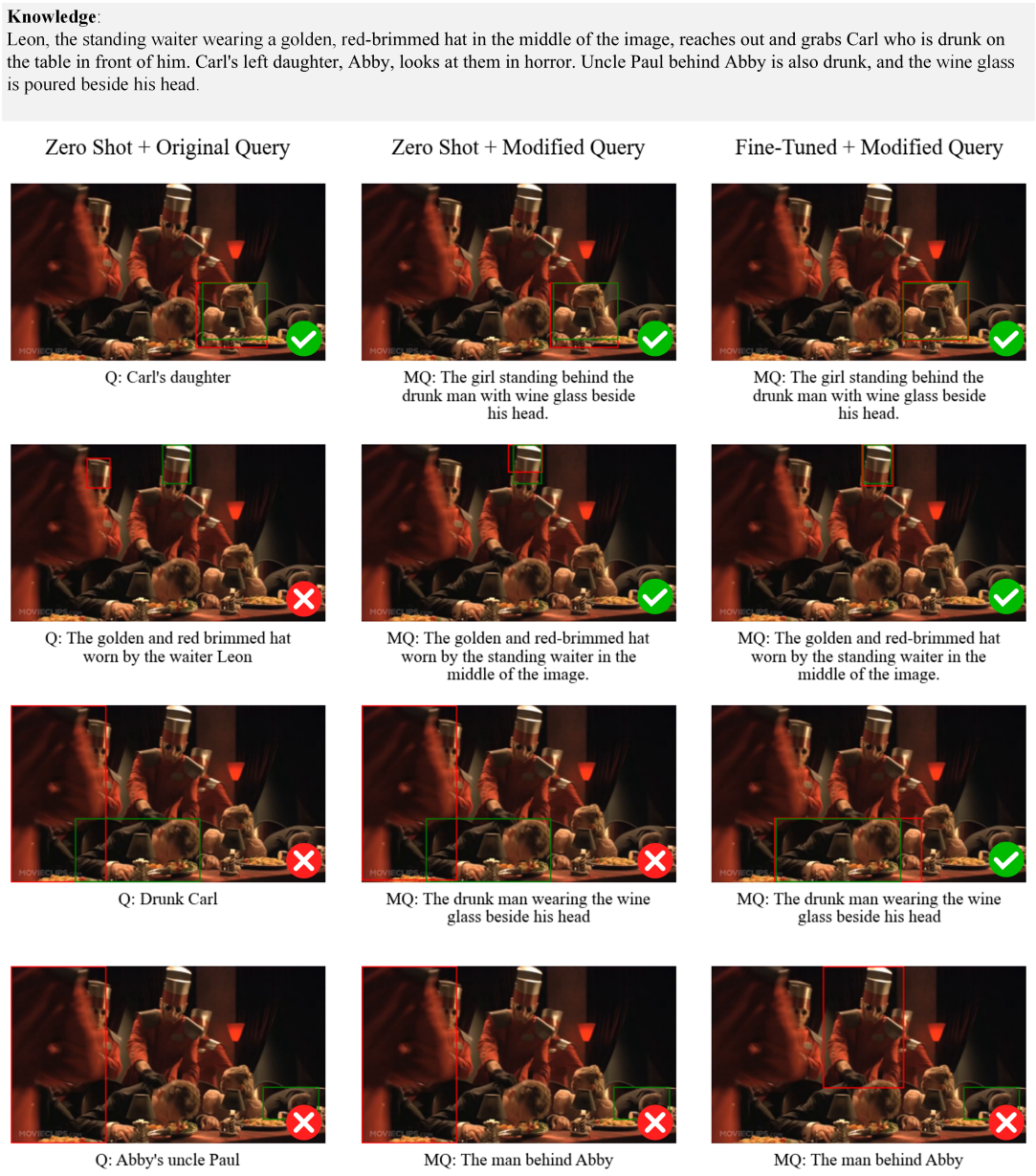



Figure 5.2: Qualitative Analysis on VG Performance in Various Examples. We compared OFA Large Performance on 3 evaluation settings: 1. Zero shot with original query (left column), 2. Zero shot with modified query (middle column), and 3. Fine Tuned version with modified query.

tendency for Llama 3 to transform it so that the main focus becoming the person. As example, in row 1, the original query ”the black top hat worn by Leon” which focuses on ”hat” is rewritten as ”the man in black top hat”, shifting the focus to ”the man”. Additionally, there are some other samples that where query modification alters the meaning of the expression, illustrated in Figure 5.5.

We also investigate how the parameter size affects the performance of the VG models. Some VG results with modified query of $\text{OFA}_{\text{Base}}$, $\text{OFA}_{\text{Large}}$, and $\text{OFA}_{\text{Huge}}$ is shown in Figure 5.6. In the second row, unlike $\text{OFA}_{\text{Large}}$ and $\text{OFA}_{\text{Huge}}$, $\text{OFA}_{\text{Base}}$ misfocuses the main object referred in the expression. In expression ”The colleague of Jakson who wears a striped tie”, the focus should be on the person, not the striped tie. A similar issue is observed in the third row, where $\text{OFA}_{\text{Base}}$ and $\text{OFA}_{\text{Large}}$ focus on the necklace, instead of the person (”the wife”), which $\text{OFA}_{\text{Huge}}$ correctly identifies. These focus phenomenon may explain the performance improvement seen in LeViLM[4] when linguistic structure is provided (see Table 5.1 ID: FT2 and FT3). By including linguistic structure, the model has information to know what is the head or main noun of a phrase, thereby enabling it to focus on the correct target when choosing thee bounding box.

The size of the bounding box target also impacts the system performance. In the fourth row, based on the IoU score, all model results are classified as incorrect. However, a closer look suggests that the model likely actually know the right region referred by the expression (”the watch”). The issues arises because the bounding box representation in the system is built by transforming continuous coordinates into discrete location token cell, which reduce granularity. Consequently, the model struggles in some cases involving small bounding box that require more precise localization.

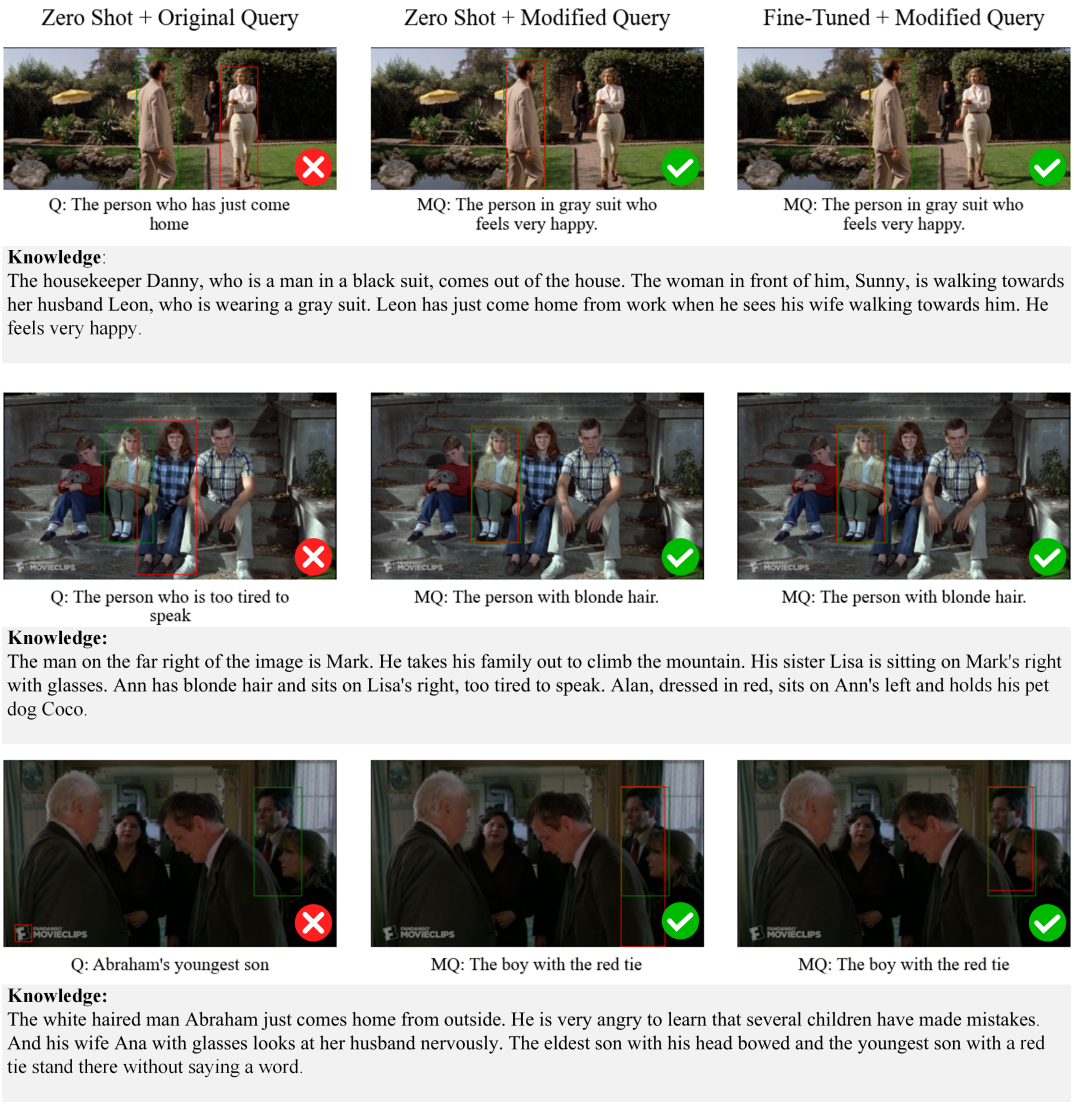



Figure 5.3: The query modification from abstract and hard to imaged visually phrases into visual description such as the clothing or physical attributes make it much easier for VG model to interpret.

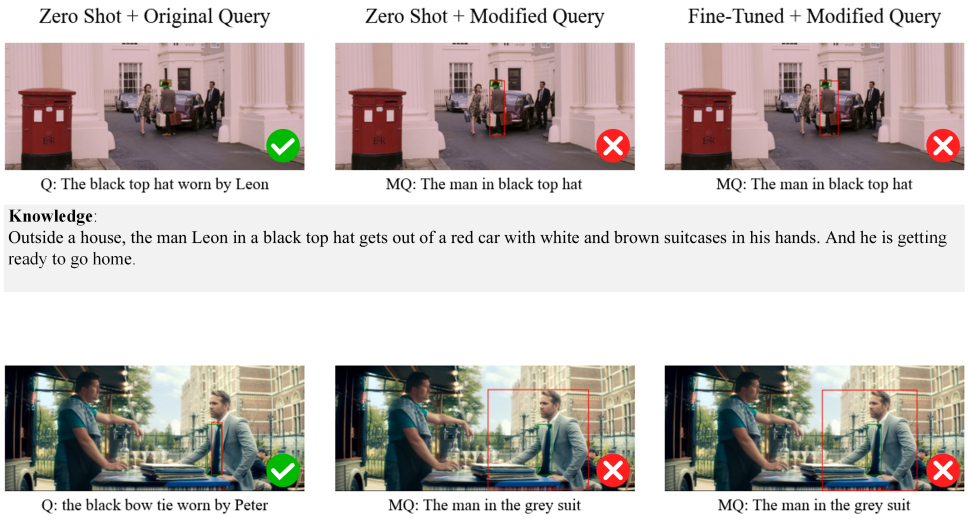



Figure 5.4: We notice that there is a tendency for LLM to shift the focus from an object to the person.

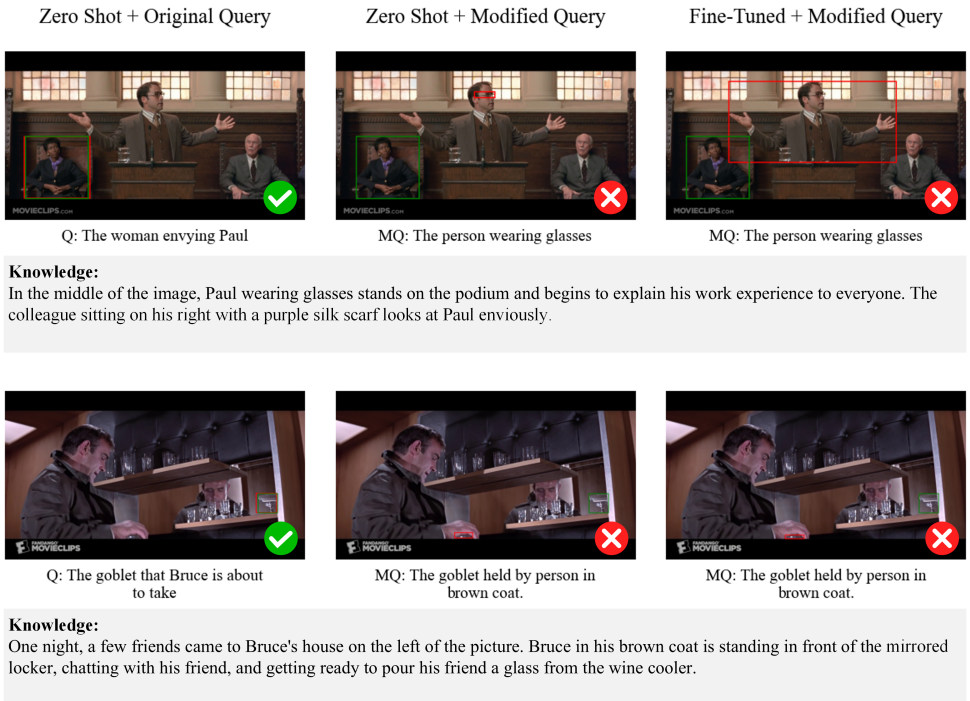



Figure 5.5: Some samples in easy category that were initially accurate in predicting bounding box with original query, becoming incorrect after query modification due to errors in the rewriting process.

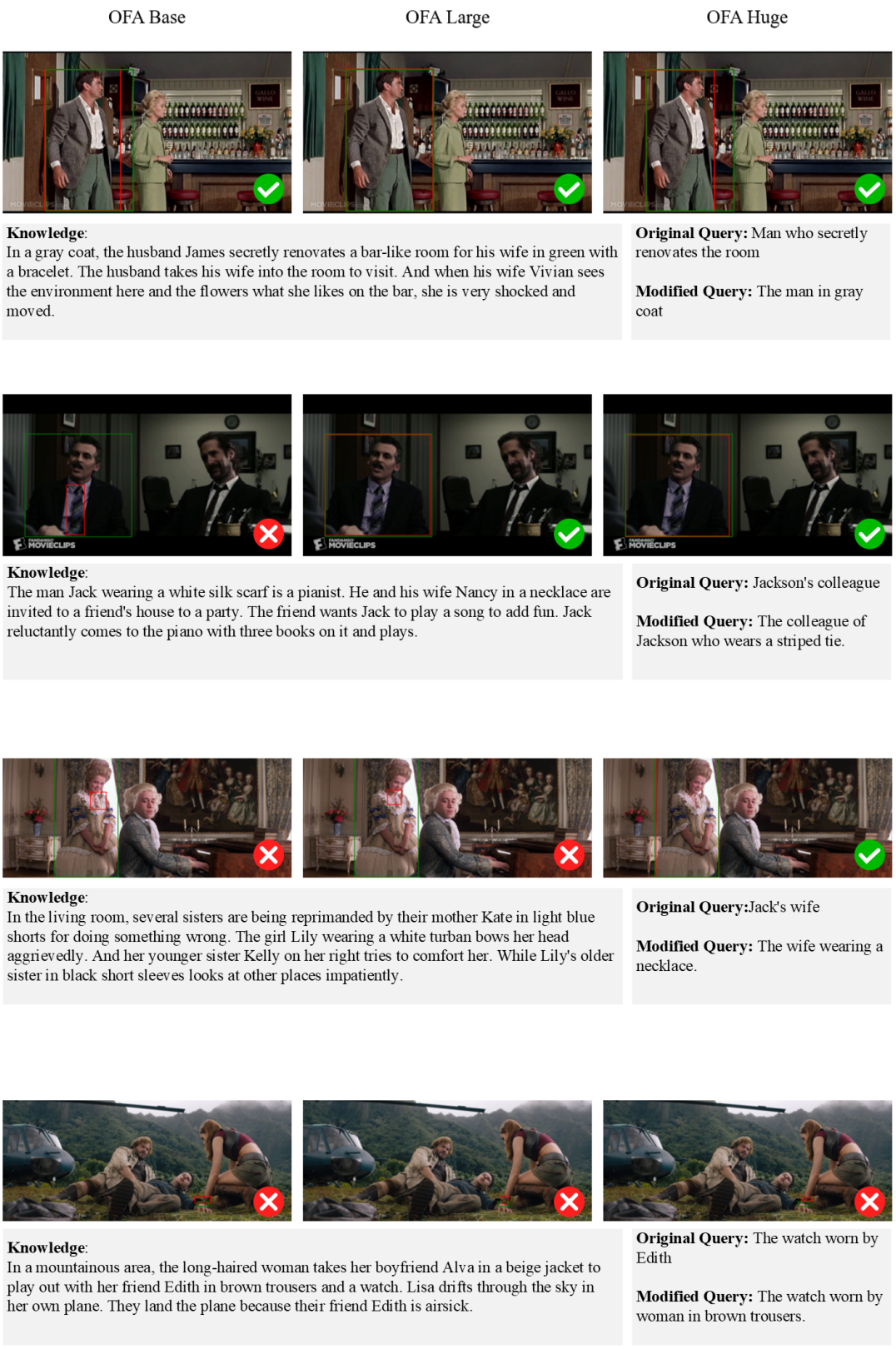



Figure 5.6: VG Performance of Various Model with Different Parameter Size. We compare the results of $\text{OFA}_{\text{Base}}$, $\text{OFA}_{\text{Large}}$, and $\text{OFA}_{\text{Huge}}$. Row 2-3 shows the misfocus problem of model in identifying the main noun of a phrase. In row 4, we observe that the model struggles with small bounding box.

# Chapter 6

# Conclusion and Future Works

## 6.1 Conclusion

Our work shows that pretrained visual language models (VLMs), particularly Kosmos-2 and OFA has demonstrated capability in solving visual grounding task with reasoning requirement. In zero-shot setting with query only, Kosmos-2, $\text{OFA}_{\text{Large}}$ and $\text{OFA}_{\text{Huge}}$ outperform LeViLM, the previous existing method. Our study also indicates that the involvement of Llama 3 to modify the query noticeably improves model accuracy. The reasoning ability of Llama 3 successfully revises the query to replace the proper noun with with visual descriptions or spatial relationship, making it much easier for grounding model to accurately detect the bounding box. When building the prompt in LLama 3 for query modification, our findings confirm that providing examples inside prompt (in-context learning) boost the quality of rewritten query, aligning with previous research on LLM in-context learning. As a result, it positively affects the performance of visual grounding model at the end. Conversely, there are some examples where an originally correct prediction becomes inaccurate after query modification. This issue roots from Llama 3 reasoning inaccuracy, such as the tendency to shift the focus from object to the person or completely alter the meaning of the original expression.

Compared to zero-shot, fine-tuning setting results in higher accuracy, as the model learns the specific characteristic of images in the SKVG dataset (which consists of movie scene) and the alignment of the bounding box and expressions. In comparison with the fine-tuned version of existing approach(LeViLM), our fine tuned $\text{OFA}_{\text{Huge}}$ achieves competitive result in the setting involving the query and knowledge only. It is crucial to note that OFA is a generalized model, contrast with LeViLM, which is specifically designed for visual grounding tasks. However, overall, the fine-tuned

version of LeViLM that incorporates linguistic structure (Q+K+S version) still reach the highest accuracy. Interestingly, our fine tuned $\text{OFA}_{\text{Huge}}$ show its superiority in hard category. The absence of linguistic structure in our model may contribute to its misfocusing issues while choosing the bounding box. We observe that in some examples when the expression involves a person with object attributes, such as ”the man who wears a striped tie”, the grounding model incorrectly choose the object (”tie”) instead of focusing on the person (”the man”). Including the linguistic structure would provide the model with the information needed to determine the main noun of a phrase, enabling it to focus on the correct target. One possible method is to add a dependency parsing pipeline and integrate its results as part of the text input of the OFA transformer.

## 6.2 Future Work

There is still some room for improvement and further research opportunities in this project as visual grounding with reasoning requirements is a relatively unexplored area. First, the modified query results generated by the LLM could be enhanced through prompt engineering or by experimenting with different LLMs. There are many pretrained models such as QWEN-VL[56], GroundingDINO[57], Florence 2[58] and GPT-4[13] have demonstrated noticable performance and worth exploring. For the grounding aspect, more extensive hyperparamater tuning could be explored, as this is not the main focus of the current work. In our fine-tuning effort where we try to investigate pretrained model ability in reasoning by providing knowledge and query, we only tried one instruction format. Future research could experiment with other instruction formats. Additionally, experimenting with Low Rank Adaption (LoRA)[59] instead of full fine-tuning could be studied. Our findings also observe that, in some example, OFA struggles in correctly choosing the main focus of the phrase. Experimenting in fine-tuning involving linguistic structure might address this issue and improve the accuracy. Last, the current prediction process still lacks of interpretability, which could be an interesting topic for further research.

# Bibliography


[1] Hang Su, Wen Qi, Jiahao Chen, Chenguang Yang, Juan Sandoval, and Med Amine Laribi. Recent advancements in multimodal human–robot interaction. *Frontiers in Neurorobotics*, 17:1084000, 2023.

[2] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. A survey on multimodal large language models for autonomous driving, 2023.

[3] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara 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 14*, pages 69–85. Springer, 2016.

[4] Yibing Song, Ruifei Zhang, Zhihong Chen, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15039–15049, 2023.

[5] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*, 2018.

[6] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10591–10599, 2019.

[7] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-language pre-training: Basics, recent advances, and future trends, 2022.

[8] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[10] 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.

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[12] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.

[13] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey,

Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu,

Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[16] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton,

Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[20] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter

Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada

Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vi-

jai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[21] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.

[22] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild, 2020.

[23] Ross Girshick. Fast r-cnn, 2015.

[24] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022.

[27] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[29] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual

concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[30] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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 (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[31] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[32] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[34] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 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, 2017.

[35] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[36] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions, 2016.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.

[38] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[39] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[40] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[42] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[43] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean,

Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[44] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[45] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models, 2023.

[46] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. *Coyo-700m: Image-text pair dataset*, 2022.

[47] 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.

[48] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces, 2022.

[49] Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. Torchscale: Transformers at scale, 2022.

[50] Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Foundation transformers, 2022.

[51] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer, 2022.

[52] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[55] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning, 2019.

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

[57] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024.

[58] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks, 2023.

[59] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

# Appendix A

# Fine-Tuning Hyperparameter

| Hyperparameter | Value |
|---|---|
| label smoothing | 0.1 |
| learning rate | 3e-5 |
| warmup ratio | 0.06 |
| batch size | 4 |
| update freq | 8 |
| resnet drop path rate | 0.0 |
| encoder drop path rate | 0.1 |
| decoder drop path rate | 0.1 |
| attention dropout | 0.0 |
| max source length | 80 |
| max target length | 20 |
| num bins | 1000 |
| patch image size | 512 |

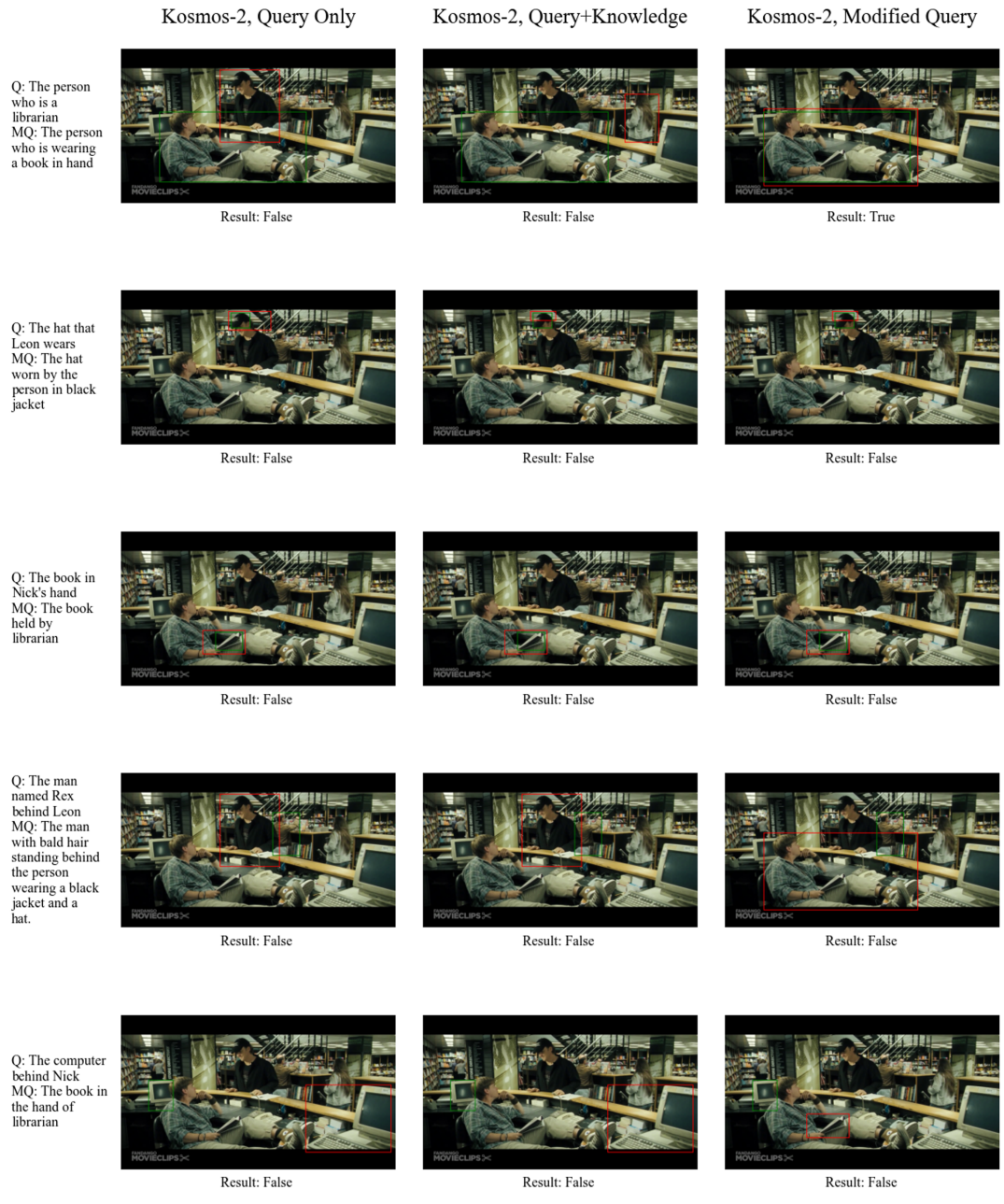
Table A.1: Hyperparameter for OFA fine-tuning.

# Appendix B

# More Visual Grounding Results

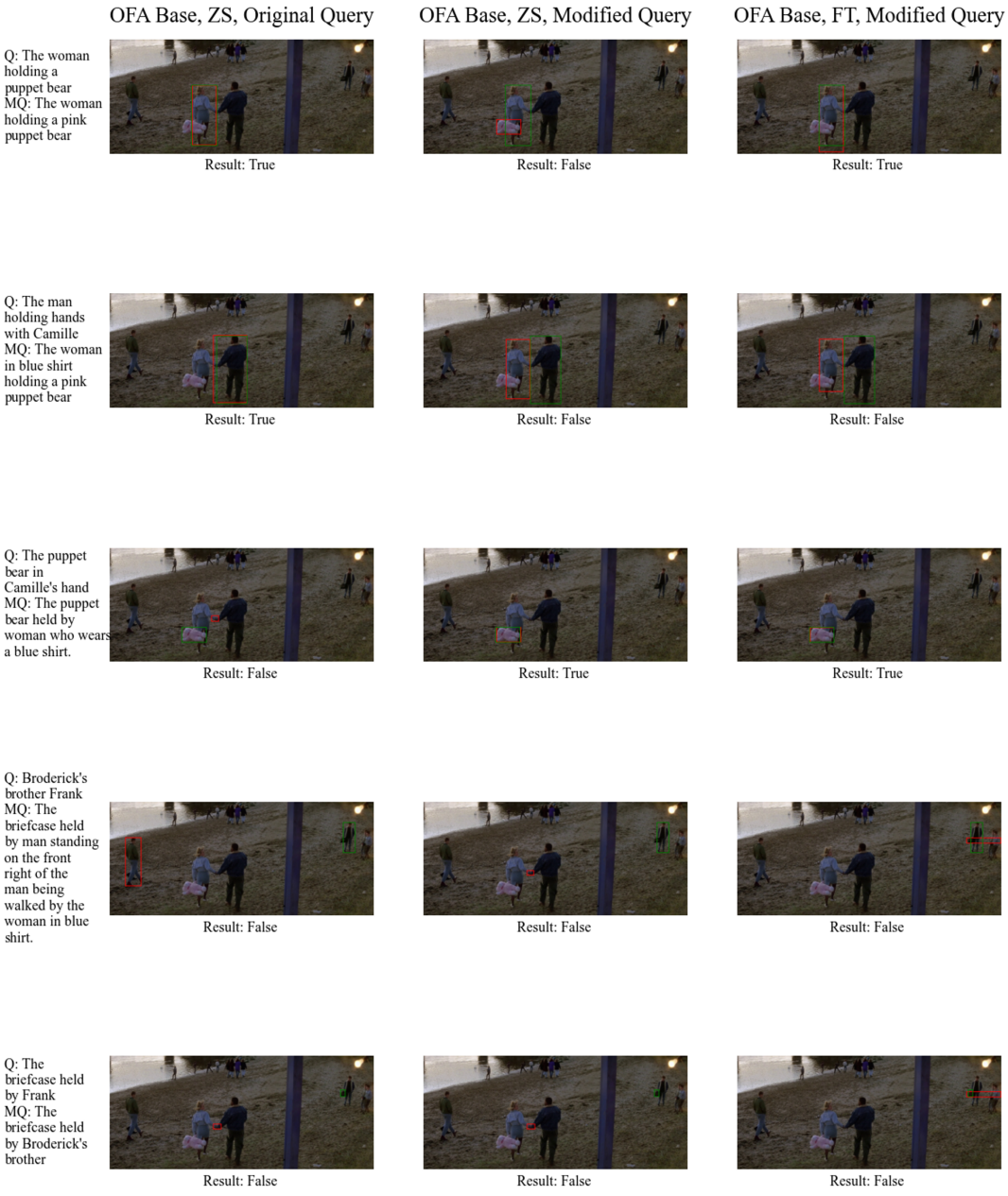

**Knowledge:**
The housekeeper Danny, who is a man in a black suit, comes out of the house. The woman in front of him, Sunny, is walking towards her husband Leon, who is wearing a gray suit. Leon has just come home from work when he sees his wife walking towards him. He feels very happy.'

Figure B.1: Examples of Kosmos-2 VG Result in Zero Shot



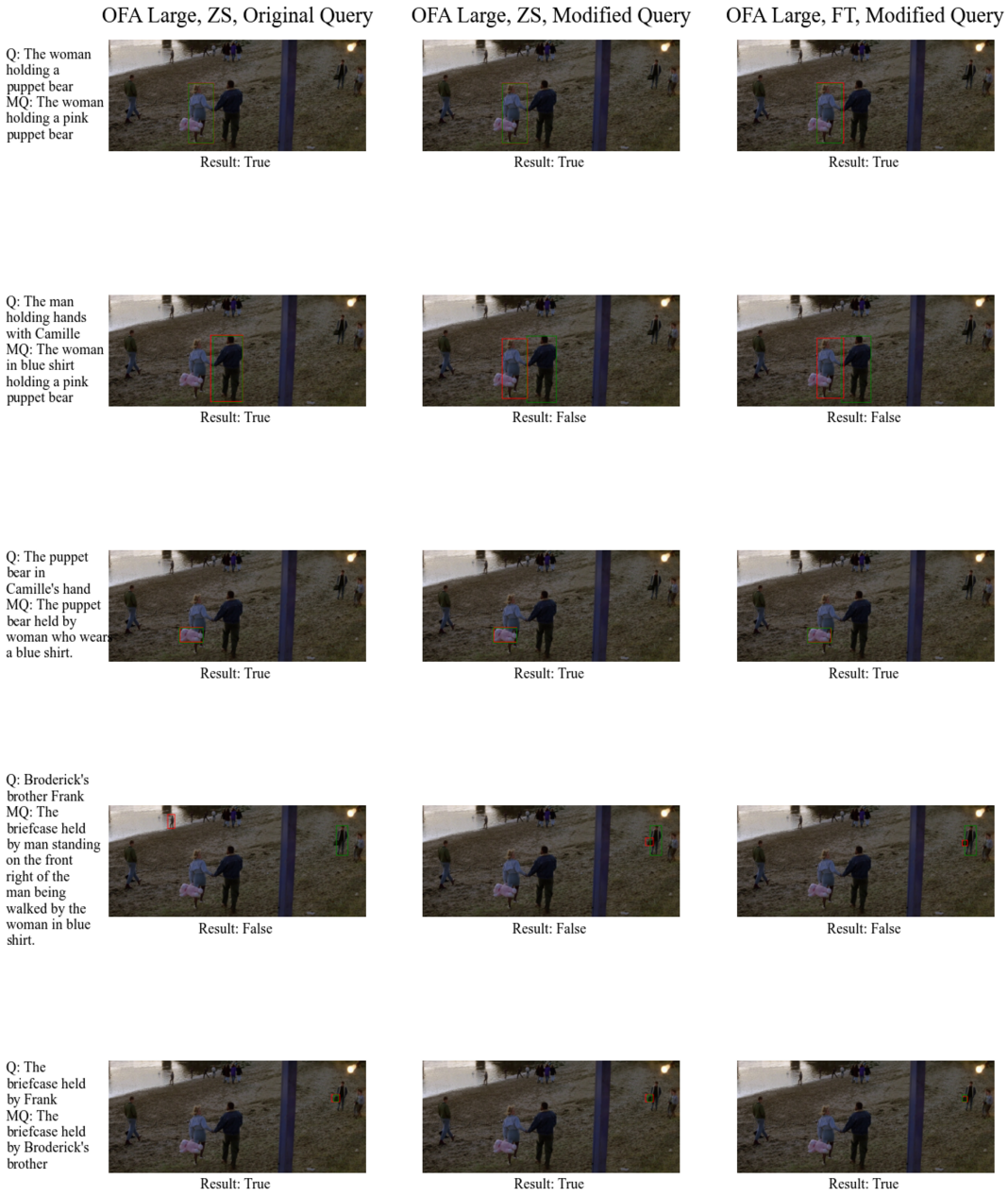

**Knowledge:**
Leon comes to the library to borrow books. He wears a black jacket and a hat, standing on the middle of the image. He is consulting Nick, the librarian who is sitting in front of him at the counter with a book in his hand. The bald man Rex is standing behind Leon and reading a book on the shelf.

Figure B.2: Examples of Kosmos-2 VG Result in Zero Shot



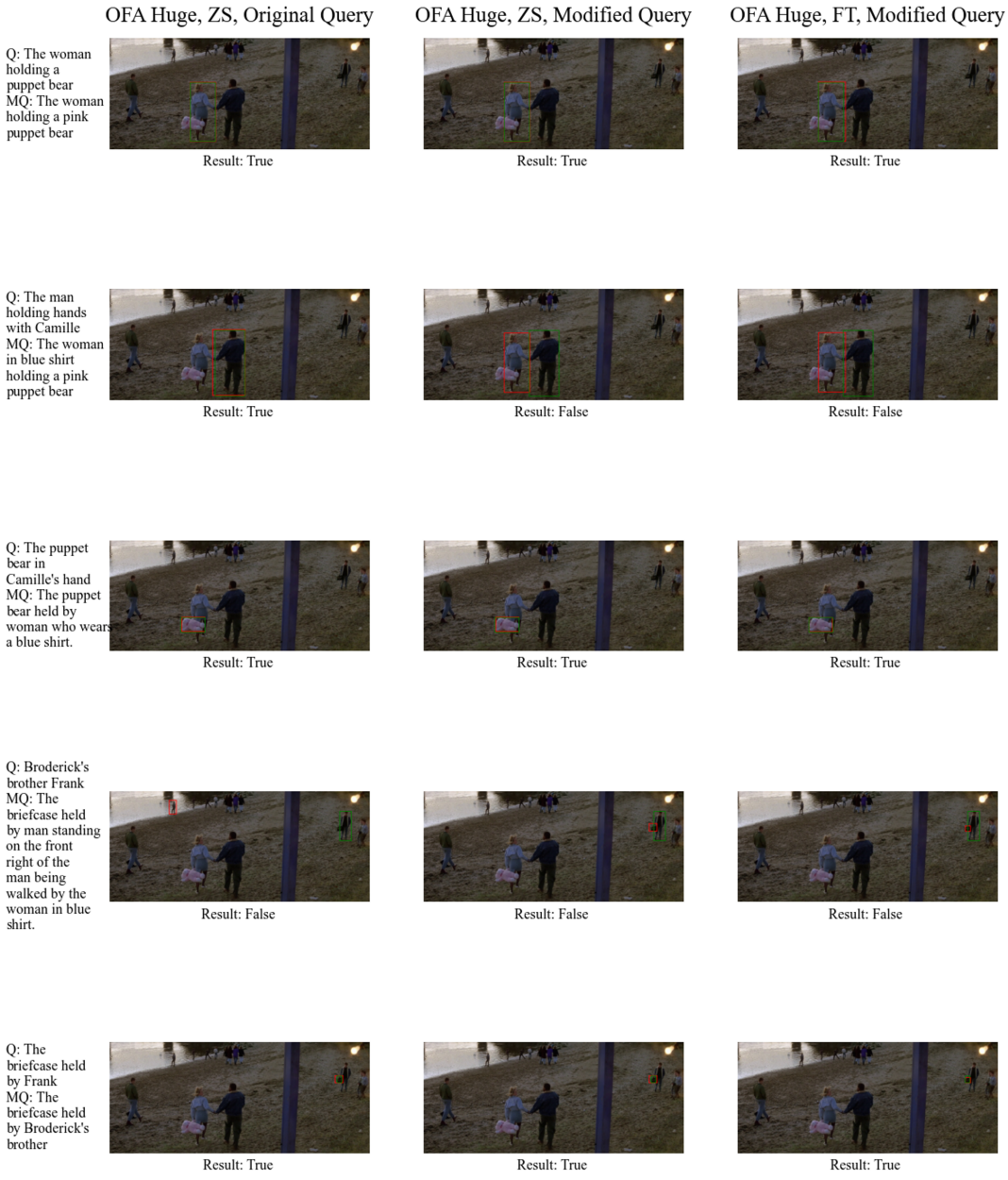

**Knowledge:**
The woman Camille wears a blue shirt and holds a pink puppet bear in her hand. She is taking her boyfriend Broderick's hand for a walk. Broderick's brother Frank is standing on the front right of Broderick, holding a briefcase and waiting for his brother.

Figure B.3: VG Results with $\text{OFA}_{\text{Base}}$




**Knowledge:**
The woman Camille wears a blue shirt and holds a pink puppet bear in her hand. She is taking her boyfriend Broderick's hand for a walk. Broderick's brother Frank is standing on the front right of Broderick, holding a briefcase and waiting for his brother.

Figure B.4: VG Results with $\text{OFA}_{\text{Large}}$




**Knowledge:**
The woman Camille wears a blue shirt and holds a pink puppet bear in her hand. She is taking her boyfriend Broderick's hand for a walk. Broderick's brother Frank is standing on the front right of Broderick, holding a briefcase and waiting for his brother.

Figure B.5: VG Results with $\text{OFA}_{\text{Huge}}$