ParGo: Bridging Vision-Language with Partial and Global Views

An-Lan Wang[1, 2, †, *], **Bin Shan**[1, *], **Wei Shi**[1], **Kun-Yu Lin**[2], **Xiang Fei**[1], **Guozhi Tang**[1], **Lei Liao**[1],
Jingqun Tang[1], **Can Huang**[1], **Wei-Shi Zheng**[2]

[1]ByteDance China
[2]School of Computer Science and Engineering, Sun Yat-sen University, China
{wanganlan, shanbin, shiwei.11, feixiang.77, tangguozhi.997, liaolei.666, can.huang}@bytedance.com,
linky5@mail2.sysu.edu.cn, wszheng@ieee.org

Abstract

This work presents ParGo, a novel Partial-Global projector designed to connect the vision and language modalities for Multimodal Large Language Models (MLLMs). Unlike previous works that rely on global attention-based projectors, our ParGo bridges the representation gap between the separately pre-trained vision encoders and the LLMs by integrating global and partial views, which alleviates the overemphasis on prominent regions. To facilitate the effective training of ParGo, we collect a large-scale detail-captioned image-text dataset named ParGoCap-1M-PT, consisting of 1 million images paired with high-quality captions. Extensive experiments on several MLLM benchmarks demonstrate the effectiveness of our ParGo, highlighting its superiority in aligning vision and language modalities. Compared to conventional Q-Former projector, our ParGo achieves an improvement of 259.96 in MME benchmark. Furthermore, our experiments reveal that ParGo significantly outperforms other projectors, particularly in tasks that emphasize detail perception ability. The code and models are released at https://github.com/AlanWang0o0/ParGo

Introduction

Recent Multi-Modal Large Language Models (MLLMs) (OpenAI 2023b; Team et al. 2023; Liu et al. 2023b; Li et al. 2022) achieve remarkable progress across various tasks (*e.g.*, Visual Question Answering (VQA)). The vision-language projector as a widely used component in MLLMs, aims to provide LLMs with proper visual features. Due to its critical role in bridging modalities, it has garnered significant attention in recent research (Cha et al. 2023; Alayrac et al. 2022; Zhu et al. 2023).

The pioneer works (Zhu et al. 2023; Liu et al. 2023b) directly project the visual feature using linear or Multi-Layer Perceptron layer (MLP). Nevertheless, such linear-based projector struggles to control the number of visual tokens provided to LLMs (*e.g.*, handling fine-grained features), resulting in high computational costs. Another line of works (Li et al. 2023b; Alayrac et al. 2022), employing global attention-based projectors, perform a global projection of image features to a fixed number of visual tokens using attention operation. However, these projectors based on global projection lead to the produced tokens concentrating on prominent regions while overlooking finer details. Take the image in Figure 1 as an example, previous methods tend to focus on the fortress and easily overlook the two individuals at the top.

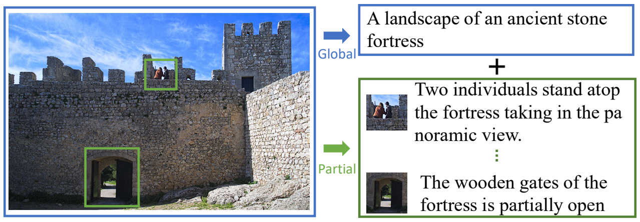

A landscape of an ancient stone fortress
Global
+
Partial
Two individuals stand atop the fortress taking in the pa noramic view.
⋮
The wooden gates of the fortress is partially open

Figure 1: Illustration of the global and partial information. An image can be properly described by the two kinds of information. Globally, this image shows a landscape of an ancient stone fortress. Delve into the partial information, two individuals stand atop the fortress, the wooden gate at the bottom of the fortress is partially open, and so forth. Please zoom in and view in color.

In this paper, we aim to build a vision-language projector that can provide the LLMs with visual features that better represent the image, while using a fixed number of visual tokens. The inspiration stems from the observation that an image can be properly described by two kinds of information, namely, global information presents a holistic understanding of images, while multiple partial information emphasizes the subtle details, an example is shown in Figure 1.

Motivated by this, we propose a novel Partial-Global projector (ParGo) based on a partial-global attention mechanism. By integrating both global and partial views, our ParGo effectively bridges the representation gap between separately pre-trained vision encoders and LLMs, alleviating the overemphasis on prominent regions. In addition, considering the relation between different partial regions in an image, ParGo incorporates a cascaded partial perception block, which enables interaction between different partial regions of an image.

Furthermore, to facilitate the effective training of ParGo, we collect a large-scale detail-captioned image-text dataset named ParGoCap-1M-PT for pre-training. Most existing pre-training datasets, typically sourced from the Internet,

* Equal contribution. † This work is done when An-Lan Wang is an intern at ByteDance China.

contain captions that are usually short and emphasize prominent visual features while lacking detailed descriptions of partial regions. Training on such datasets makes it challenging for the model to learn fine-grained details. In contrast, our ParGoCap-1M-PT contains longer and more detailed descriptions of multiple regions in images. Pre-trained on the two kinds of captioned data, we transfer our models into multiple downstream tasks using several public-available instruction tuning datasets, *e.g.*, LLaVA-150k (Liu et al. 2023a). Extensive experiments are conducted on several benchmarks, and the results demonstrate that our Partial-Global projector outperforms other projectors. Our contribution can be summarized:

- We propose a novel Partial-Global projector (ParGo) that well aligns two separately pre-trained models by integrating partial and global views, alleviating the overemphasis on prominent regions.
- To facilitate the modality alignment, we further propose a new detail-captioned pre-training dataset, ParGoCap-1M-PT, including 1 million images paired with high-quality captions.
- Extensive experiments on several MLLM benchmarks demonstrate the effectiveness of our proposed ParGo. Our projector achieves state-of-the-art results compared with previous projectors, particularly excelling in some tasks that need more detail-perception ability.

Related Work

Multi-modal Large Language Models

Large Language Models (LLMs) (Touvron et al. 2023; OpenAI 2023b; Chiang et al. 2023) have achieved remarkable progress, leading recent works (Alayrac et al. 2022; Team et al. 2023; Bai et al. 2023) to generalize this success to more modalities, *i.e.,*, Multimodal Large Language Models (MLLMs). In those works, closed-source works (OpenAI 2023b; Team et al. 2023; Bai et al. 2023) have shown great advancements, highlighting their high performance on complex tasks. In contrast, open-source models have also made significant progress, promoting transparency and collaboration in the research community. Pioneer works (Li et al. 2023b; Alayrac et al. 2022) established competitive baselines by integrating massive image-text pairs. Furthermore, recent works (Liu et al. 2023b; Dai et al. 2023) (Liu et al. 2023a,b; Chen et al. 2023b) boosting the zero-shot capabilities on various downstream tasks via collecting more high-quality multi-modal instruction (visual instruction tuning). On the other hand, recent works (Liu et al. 2024a; Hu et al. 2023; Liu et al. 2024b) also focus on fine-grained understanding (*e.g.*, text recognition).

Vision-language Projector

Vision-language projectors play a crucial role and are widely used components in MLLM. They aim to connect the visual feature space and language feature space, which can be divided into linear-based and attention-based projectors. Linear-based projectors (Liu et al. 2023b,a; Zhu et al. 2023; Chen et al. 2023b; Dong et al. 2024) employ a linear layer to connect the vision encoder seamlessly with the language model (LLM). Despite their straightforward implementation, the linear-based projectors encounter challenges in producing a large number of visual tokens to LLMs, leading to high computational costs. Another line of research (Alayrac et al. 2022; Li et al. 2023b; Bai et al. 2023; Dai et al. 2023; Ye et al. 2023b) explore more flexible projectors (*e.g.*, Q-former (Li et al. 2022) and Perceiver Resampler (Alayrac et al. 2022)) based on attention mechanism. Such attention-based methods often extract prominent image features, leading to a loss of detail and a drop in the model performance. Similar findings are also mentioned in a recent work, Honeybee (Cha et al. 2023), which proposes a D-Abstractor that uses a Deformable attention (Zhu et al. 2020) to retain the local information and achieve superior performance. To efficiently provide comprehensive information to LLMs using a fixed number of visual tokens, we propose Partial-Global projector, which uses a partial-global projection that simultaneously extracts both partial and global information.

Multi-modal Pre-training Data

Training models using web-crawled large-scale image-text datasets (*e.g.*, (Schuhmann et al. 2021; Byeon et al. 2022; Changpinyo et al. 2021; Sharma et al. 2018)) has become the most common strategy for MLLMs. Nevertheless, web-crawled datasets primarily present the main feature of the image using noisy and short captions, lacking detailed descriptions. To obtain detailed descriptions, some works provide boxes (or mask) level captions but are constrained by the box-generation model. The recent remarkable progress achieved by close-sourced MLLMs (OpenAI 2023b; Team et al. 2023) has led recent researchers(Chen et al. 2023b; Yu et al. 2024a) to consider using MLLM to synthesize detail-captioned data, supplementing the limitations of conventional web-crawled datasets. In this work, we further contribute a detail-captioned dataset for pre-training, aimed at enhancing the alignment between the two modalities from a data perspective.

Methodology

Overview

In Figure 2, (a), we illustrate the pipeline of the MLLM that uses our proposed Partial-Global projector (ParGo) as the vision-language projector. Given an image I and related text T, we first use a frozen image encoder and a tokenizer to extract the visual feature f_v and text feature f_t, respectively. To effectively align the vision and language modalities, we propose a Partial-Global projector to project the visual feature to the text feature space. Specifically, the Partial-Global projector projects the visual features from the partial and global views, employing two kinds of learnable tokens. Subsequently, the outputs of ParGo and the tokenized input text are fed into the Large Language Model to generate the final text output.

Partial-Global Projector

To better align the separately pre-trained visual encoder and large language models, we propose the Partial-Global pro-

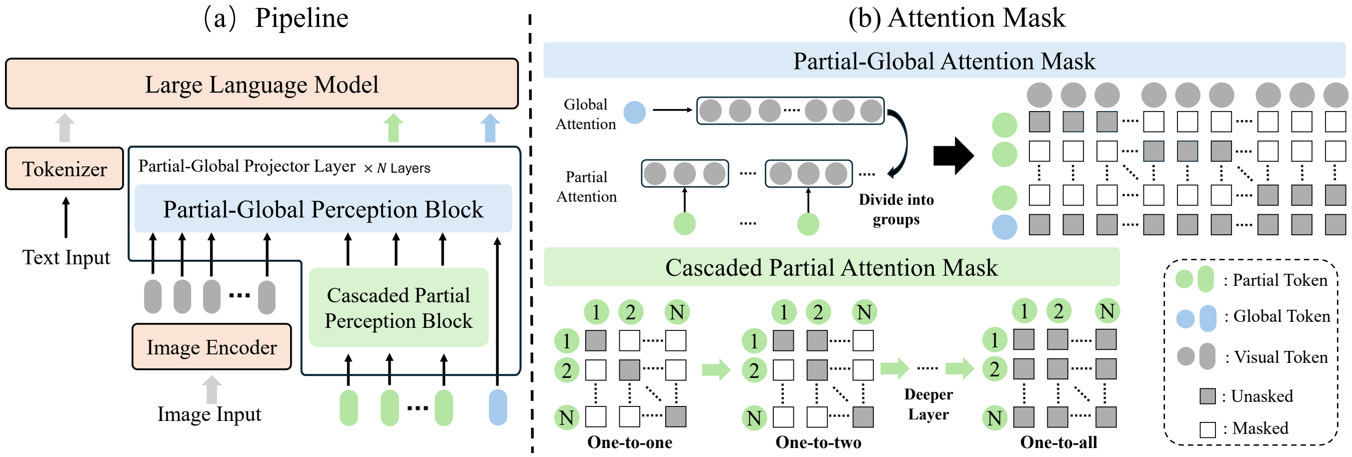


Figure 2: (a). The pipeline of a MLLM with our proposed ParGo as the vision-language projector. First of all, we use a frozen image encoder to extract image features. To better align the pre-trained visual encoder with the LLM, we propose a Partial-Global projector to project the image features using two kinds of tokens *i.e.,* partial and global tokens. Finally, the output partial and global visual tokens, as well as the tokenized text, are fed into the LLM to generate the text output in an auto-regressive manner. Specifically, each Partial-Global projector layer contains a Partial-Global Perception block that utilizes two kinds of tokens, *i.e.*, partial and global tokens, to extract the image features. Additionally, to fully consider the relation between different partial regions in an image, a cascaded partial perception block is incorporated to enable interactions between partial tokens in a cascaded manner. (b). A Demonstration of the Partial-Global and the Cascaded Partial Attention mask. It's worth noting that the Partial-Global Attention mask remains the same in different layers, while the Cascaded Partial Attention mask changes across various layers.

jector (ParGo). Each ParGo layer mainly contains two parts, *i.e.*, a Partial-Global Perception block and a Cascaded Partial Perception block, as illustrated sequentially,.

Partial-Global Perception Block. Firstly, we propose a Partial-Global Perception block, which employs two kinds of tokens, *i.e.,* partial tokens and global tokens, to extract the partial and global information, respectively.

In detail, given the visual features $f_v \in \mathbb{R}^{n_v \times c}$ (extracted by the visual encoder), we randomly initialize a set number of global tokens $q_g \in \mathbb{R}^{n_g \times c}$ and partial tokens $q_p \in \mathbb{R}^{n_p \times c}$, where c is the feature dimension, and n_v, n_g, n_p is the number of visual features, global tokens, and partial tokens, respectively. The tokens interact with the image features in a cross-attention layer. For the cross-attention mask, we use a pre-defined mask $M \in \mathbb{R}^{(n_p+n_g)\times n_v}$, which ensures that one partial token only interacts with part of the visual features, while the global tokens interact with all visual features. The number of visual features each partial token interact with n_s is calculated as follows:

$$n_s = n_v/n_p, \tag{1}$$

An example of the mask used in the Partial-Global Perception block is shown in Figure 2.(b). Intuitively, each partial token only *sees* part of the image and global tokens *see* the whole image.

Cascaded Partial Perception Block. In the above partial-global attention process, the partial tokens *see* consistent visual tokens in different layers. However, such a process focuses on using one token to project part of the image, thus it may be unable to fully consider the relation between different partial regions in an image. Therefore, to fully consider the context of the partial image, we propose a Cascaded Partial Perception (CPP) block that enables interaction between different partial tokens.

Specifically, in addition to the Partial-Global Perception block, we insert a Cascaded Partial Perception (CPP) block in front of it. The CPP block is implemented using a masked self-attention block with a specially designed mask as the core. The input is the partial tokens $\{q_p^i\}_{i=1}^{n_p}$, and as the depth l increases, the number of adjacent tokens each partial token can *see* $n^l_{vis}(0 \leq n^l_{vis} \leq n_p)$ increase linearly as well. This process can be formulated as follows:

$$\{n^{\,l}_{vis} = k \times l, k = n_p/d \tag{2}$$

where l is the index of the layer, n_p is the number of partial tokens, and d is the number of the Partial-Global projector layer. As shown in Figure 2. (b), an example of the mask in different CPP blocks in different layers is visualized.

ParGoCap-1M-PT

In this section, we focus on how to pre-train the overall model to leverage the advantage of the Partial-Global projector. Existing pre-training corpora are generally coarse-captioned, and collected from the Internet. Such captioned data are noisy and are usually short (average character number less than 100), describing only part of the image. Using such coarse data hinders the alignment between the two modalities. Therefore, we construct a new large-scale detail-captioned image-text dataset, named **ParGoCap-1M-PT**, using off-the-shell closed-source MLLMs. Table 1 compares existing caption datasets and our ParGoCap-1M-PT.

ParGoCap-1M-PT offers a large amount of high-quality detailed captioned samples, which distinguishes itself from existing datasets. The data collection pipeline mainly consists of two steps:

Detailed caption generation. To facilitate the alignment between the vision and language feature space, large-scale and diversified image data is required. We first randomly select a large number of images from the Laion dataset (Schuhmann et al. 2021). Then, to generate detailed captions that well describe the image, we employ the powerful close-sourced MLLMs (*i.e.,* GPT4-V and Gemini) to generate the captions given a specified prompt. Since our goal is to generate captioned data that takes into account both partial and global information about an image, we design the prompt that asks the MLLMs to describe the image globally and partially.

Quality control. Thanks to the powerful capabilities of existing models (OpenAI 2023a; Team et al. 2023), the quality of the generated data is already very excellent. However, there may still be some erroneous data due to the hallucination problem. To further filter out high-quality data, we employed a simple but effective quality control method. Following previous works (Li et al. 2022), we directly use several models (Radford et al. 2021; Li et al. 2022) to calculate the similarity between the image and generated captions. Image-caption pairs with low similarity are dropped. This step filters out a small portion of the data, proving that our data quality is exceptional. For more dataset details, please refer to the supplementary materials.

Experiment

Benchmarks and Metrics

To thoroughly validate the superiority of the proposed ParGo, we utilize four benchmarks tailored specifically for evaluating Multi-modal Large Language Models (MLLMs), including MME (Fu et al. 2023), MMBench (Liu et al. 2023c), SEED-Bench (Li et al. 2023a), and MM-Vet (Yu et al. 2024b). The First three benchmarks evaluate a range of MLLM capabilities, including perceptual understanding and visual reasoning. They employ different question formats: MME uses binary yes/no questions, while MMBench and SEED-Bench utilize multiple-choice questions. As for the MM-Vet, which uses the GPT-4 to evaluate the open-ended outputs of MLLMs, we include this benchmark to monitor the model's performance in natural language generation.

For all benchmarks, we report the official metrics computed using official implementation.

Implementation details.

Model Configuration. We use pre-trained EVA-02-CLIP-L/14 (Sun et al. 2023) with 336 resolution as the visual encoder. For the Large Language model, we employ the 7B Vicuna (Chiang et al. 2023) for a fair comparison. For the Partial-Global projector, six layers are utilized for all experiments if not otherwise specified. The number of partial and global tokens is 288 and 16, respectively, resulting in a total of 304 tokens.

Dataset	Captioned by	Samples	Avg.
Coarse-captioned			
COCO-Caption (Chen et al. 2015)	Human	118K	52
BLIP-LCS (Li et al. 2022)	BLIP	558K	54
LLaVA-23K (Liu et al. 2023b)	GPT4	23K	609
Detail-captioned			
ShareGPT4V (Chen et al. 2023b)	GPT4-V	0.1M	942
ShareGPT4V-PT (Chen et al. 2023b)	Share-Captioner	1.2M	826
ParGoCap-1M-PT	**GPT4-V, Gemini**	**1M**	**921**

Table 1: Comparison of existing caption datasets and our ParGoCap-1M-PT. Avg. represents the average character number of the caption.

Training. The proposed ParGo is trained using a two-stage pipeline, *i.e.*, coarse-detailed captioned pre-training and supervised fine-tuning. In the pre-train stage, we freeze the visual encoder and the LLM, focusing on training the Partial-Global projector, to gain better alignment between the two modalities. In the supervised fine-tuning stage, the visual encoder is kept frozen, and we fine-tune the Partial-Global projector and the LLM. It is worth noting that instead of training the entire LLM, we use parameter-efficient fine-tuning, *i.e.*, Low-Rank Adaptation(LoRA) (Hu et al. 2021), and the rank is set to 256. For all experiments, 32 A100 80GB GPUs are used. We employ deepspeed zero-2 (Rajbhandari et al. 2020) and flash-attention v2 (Dao 2023) for all experiments, The pre-training stage takes approximately 24 hours, while the supervised fine-tuning tasks require around 12 hours. For the ablation studies, we employ half of the training schedule (50k pre-training, 1 epoch supervised fine-tuning) compared to the final model. For more details, please refer to the supplementary materials.

Data. For pre-training, we use existing coarse-captioned data, including CC-3M, SBU Caption-1M, LAION-400M, and the constructed detail-captioned data ParGoCap-1M-PT.

For the supervised fine-tuning stage, following previous work (Cha et al. 2023), we employ four types of tasks, including: 1) Open-ended VQA, *i.e.* VQAv2 (Goyal et al. 2017), GQA (Hudson and Manning 2019), OCRVQA (Mishra et al. 2019), VSR (Liu, Emerson, and Collier 2023). 2) Multiple-Choice VQA, including ScienceQA (Lu et al. 2022), A-OKVQA (Schwenk et al. 2022). 3) Referring Expression Comprehension (REC), which includes RefCOCO (Kazemzadeh et al. 2014) and VG (Krishna et al. 2017). 4) Instruction tuning data, LLaVA150k (Liu et al. 2023b). For each dataset, we use the same templates as previous works (Liu et al. 2023a; Cha et al. 2023).

Main Results

In Table 2, we compare our ParGo with previous state-of-the-art methods. Firstly, we compare our model with the methods that use the attention-based projector. Compared with Honeybee (D-Abstractor), our ParGo obtains significant improvement, *e.g.*, 2.9 in MMBench and 77.6 in MME. Furthermore, compared with methods that use vanilla attention-based projector *e.g.,* Resampler or Q-former, our

Method	LLM	Projector	Res.	MMB	MME^P	MME	SEED	MM-Vet
MLLMs using Linear-based projectors								
LLaVA (v1) (Liu et al. 2023b)	LLaMA-7B	Linear	224	38.7	502.8	717.5	33.5	23.8
Shikra (Chen et al. 2023a)	Vicuna-7B	Linear	224	58.8	-	-	-	-
LLaVA-1.5 (Liu et al. 2023a)	Vicuna-7B	MLP	336	64.3	1510.7	1795.7	58.3	30.5
Honeybee (Cha et al. 2023)	Vicuna-7B	C-Abstractor	224	70.1	**1584.2**	1891.3	64.5	**34.9**
MLLMs using Attention-based projectors								
MiniGPT-4 (Zhu et al. 2023)	Vicuna-7B	Resampler	224	24.3	581.7	726.0	47.4	22.1
mPLUG-Owl (Ye et al. 2023a)	LLaMA-7B	Resampler	224	49.4	967.3	1243.4	34.0	-
InstructBLIP (Dai et al. 2024)	Vicuna-7B	Q-former	224	36.0	-	-	58.8	26.2
IDEFICS	LLaMA-7B	Flamingo	224	48.2	-	-	44.5	-
Qwen-VL (Bai et al. 2023)	Qwen-7B	Resampler	448	38.2	-	-	62.3	-
Qwen-VL-Chat (Bai et al. 2023)	Qwen-7B	Resampler	448	60.6	1487.5	1848.3	65.4	-
Honeybee (Cha et al. 2023)	Vicuna-7B	D-Abstractor	224	70.8	1544.1	1835.5	63.8	-
Ours	Vicuna-7B	ParGo	336	**73.7**	1579.86	**1913.1**	**67.3**	33.5

Table 2: **Comparison with other state-of-the-art MLLMs.** Res. indicates the image resolution. We highlight the **best results** and second-best results in bold and underline.

advantages are more apparent, *e.g.,* 37.7 in MMBench compared with InstructBLIP (Dai et al. 2023).

Additionally, we compare our model with the methods that use linear-based projectors, as shown in Table 2 top. Compared with methods using Linear or MLP projectors, the advantage of our Partial-Global projector is also significant, *i.e.* 117.4 improvement in MME compared with LLaVA-1.5(Liu et al. 2023a). Compared with a more recent work, Honeybee (C-Abstractor) (Cha et al. 2023), our ParGo outperforms in MMB, MME, and SEED, while slightly underperforming in MM-Vet. It is worth noting that honeybee (Cha et al. 2023) fine-tuning all the LLM parameters during the supervised fine-tuning state, while we use the parameter efficient fine-tuning strategy, *i.e.,* Low-Rank Adaptation (LoRA) (Hu et al. 2021). This full-finetuning strategy boosts the performance of Honeybee in the MM-Vet benchmark, which needs more natural language generation ability.

These consistent results demonstrate the superiority of our proposed Partial-Global projector in effectively bridging the representation gap between visual and language modalities.

Ablation Study

Effect of the components in Partial-Global projector. In Table 3, we ablate the main component in our proposed Partial-Global projector, *i.e.*, the Partial-Global Perception (PGP) block and Cascaded Partial Perception (CPP) block. For a fair comparison, the baseline uses a Q-Former with 304 tokens as the visual projector. As shown in the table, by adding the Partial-Global Perception block (*i.e.,* 16 global tokens and 288 partial tokens), we obtain a significant improvement over the baseline, *i.e.,* 162.73 in MME. These results suggest that our proposed Partial-Global projection effectively bridges the representation gap, boosting the overall performance. Then, by introducing the Cascaded-Partial Perception block, our model further obtains an improvement, *i.e.,* 92.23 in MME. In summary, the ablation study demonstrates the effectiveness of the proposed Partial-Global projector.

PGP	CPP	MME	MM-Vet
		1591.74	28.3
✓		1754.47	32.7
✓	✓	1851.70	33.1

Table 3: Ablations on the Partial-Global Perception (PGP) block and the Cascaded Partial Perception (CPP) block. The baseline uses a Q-Former as the projector.

Number of Tokens			MME	MM-Vet
Global	Partial	Total		
160	-	160	1561.42	30.1
-	144	144	1681.55	30.5
16	144	160	1802.37	32.8
16	288	304	1851.70	33.1

Table 4: Ablations on the number of partial and global tokens. Total indicates the sum of partial and global tokens.

Effect of the number of Global Partial tokens. In Table 4, we conduct a quantitative analysis of the number of partial and global tokens. Firstly, for a fair comparison, we conduct an experiment using only 160 global tokens, which achieve 1561.42 scores in MME. Then, by replacing 144 global tokens with 144 partial tokens (still 160 tokens in total), the model achieves a significant improvement (*i.e.,* 140.95 score in MME), which demonstrates the effectiveness of our partial tokens in preserving partial information. Finally, we scale up the number of partial tokens, using 288 partial tokens and 16 global tokens, and the model achieves a further improvement. Additionally, we conduct an experiment that uses only 144 partial tokens, which still outperforms the baseline with 160 global tokens, *i.e.,* 1681.55 *vs.*

Pre-training Data		MME	MM-Vet
Coarse	Detailed		
✓		1734.86	32.2
✓	✓	1851.70	33.1

Table 5: Ablations on the pre-training data. We use different combinations of Coarse captioned data and Detailed captioned data.

Projectors	# Tokens	MME	MM-Vet
Linear	576	1727.83	32.8
Q-Former	304	1591.74	28.3
ParGo	304	1851.70	33.1

Table 6: Comparison of our Partial-Global projector (ParGo) with existing Linear and Q-former projector. # Tokens means the number of tokens the projector outputs.

Large Language Model	MME	MM-Vet
Vicuna-7B (Chiang et al. 2023)	1851.70	33.1
Llama3-8B (Dubey et al. 2024)	1866.43	35.9
InternLM2-7B (Cai et al. 2024)	1869.76	37.0

Table 7: Comparison of different Large Language Models.

Projectors	MME			MM-Vet
	CNT	OCR	EXST	REC
Linear	135.0	137.5	190	37.2
Q-Former	128.33	130.0	175	34.4
ParGo	146.66	162.5	190	37.6

Table 8: Analysis on the detail perception ability. We select 4 tasks that require detail perception ability from the MME and MM-Vet benchmark, including Count (CNT), Optical Character Recognition(OCR), and Existence (EXST) from MME, Recognition (REC) from MM-Vet Benchmark.

1561.42 in MME. These consistent results demonstrate the superiority of our proposed partial tokens in preserving the partial information.

Effect of the pre-training data. In our method, to facilitate the simultaneous learning of the partial and global information, we use two kinds of pre-training data, coarse-captioned and detail-captioned. In Table 5, we ablate the pre-training data to verify the effectiveness of this strategy. As shown in the table, by adding the detail-captioned data, the model achieves an improvement in both benchmarks, *i.e.,* 116.84 in MME and 0.9 in MM-Vet. It is worth noting that the training configurations (*e.g.,* batch size and training steps) used for these experiments are the same, ensuring that the models are trained with the same amount of data. These results demonstrate the detail-captioned data help the model align the visual and language modalities.

Analysis

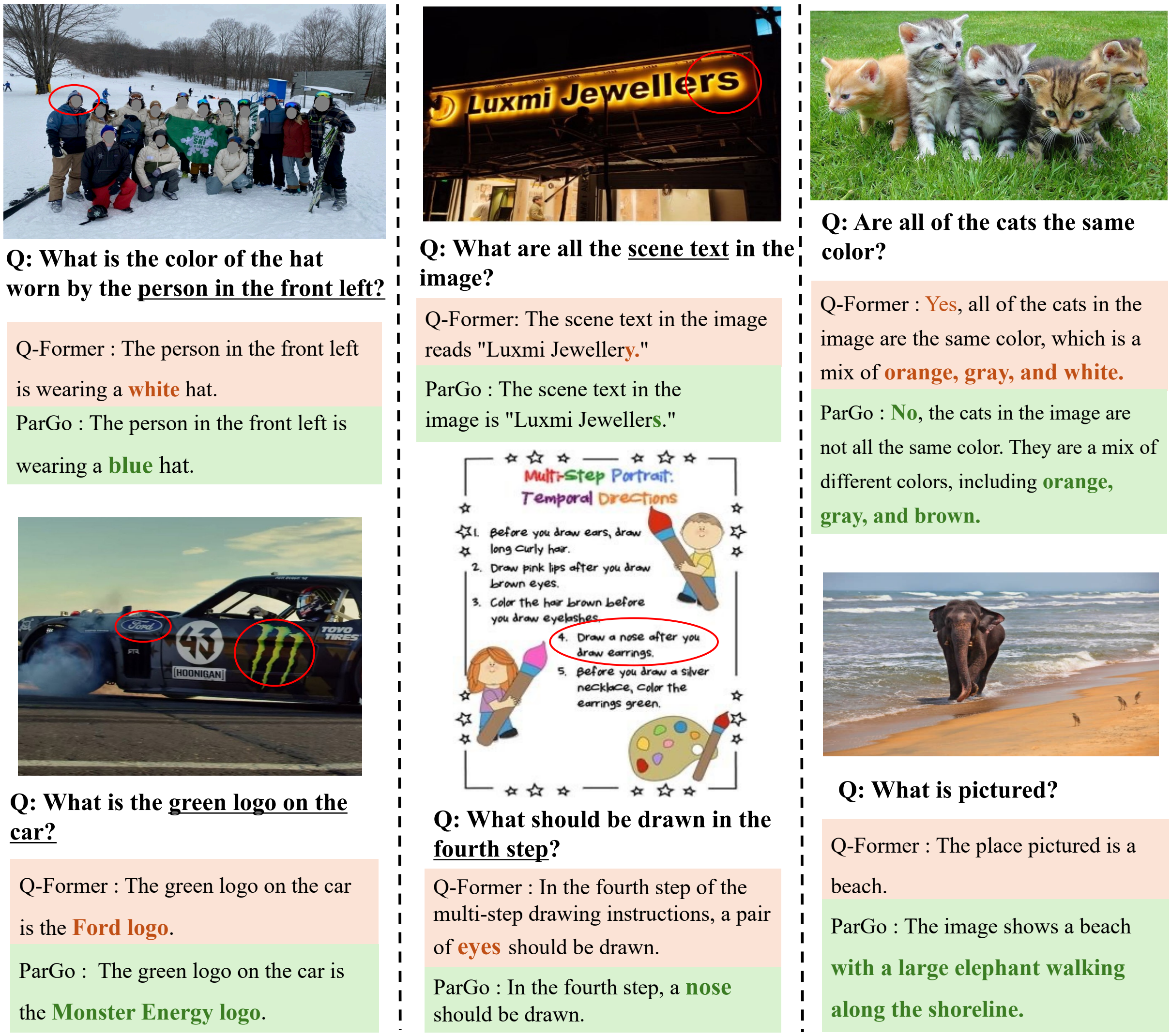
Comparison with existing visual projectors. To further illustrate the effectiveness of our proposed Partial-Global projector, in Table 6, we compare our proposed projector with a linear projector and a Q-former projector, while using the same training data and schedule for a fair comparison. The linear projector is a simple linear layer that produces 576 tokens; the Q-Former is 6-layer and produces 304 tokens. As shown in the table, the linear-based projector performs much better than the Q-Former as it uses a one-to-one projection that directly project the visual feature to the language feature space, with less information loss. However, the one-to-one projection also results in a huge number of visual tokens being fed into the LLM, introducing large computation costs. Compared with our proposed ParGo, the linear-based underperforms in both benchmarks. Our Pargo achieves an increase of 259.96 points in MME compared with the attention-based Q-Former projector. These results highlight the effectiveness of our ParGo in preserving visual information while utilizing fewer visual tokens, demonstrating its superior capability in aligning the vision and language modalities.

Comparison of different base Large Language Models. Here, we verify the generalizability of our proposed projector, we ablate several widely-used Large Language Models (LLMs), *i.e.*, Vicuna-7B (Chiang et al. 2023), Llama3-8B (Dubey et al. 2024), internLM2-7B (Cai et al. 2024). As shown in Table 7, with stronger LLM, the overall performance is further improved. Employing internLM2-7B as the base LLM achieves the best performance in both benchmarks, *i.e.,* 1869.76 in MME and 37.0 in MM-Vet. These results demonstrate the generalizability of the Partial-Global projector in aligning the visual and language modalities.

Analysis on the Partial-Global Projector. To thoroughly examine the superiority of our proposed Partial-Global projector, in Table 8, we assess the ability of different projectors in image detail perception by selecting 4 tasks that require more detail perception ability from the MME and the MM-Vet benchmarks. As shown in the table, in the tasks of OCR and CNT, which require the most detailed perception, our ParGo achieves a substantial improvement compared to the Linear projector and Q-Former Projector, *i.e.,* 16.33 in CNT and 32.5 in OCR compared with Q-Former. In the other two tasks, EXST and REC, our model also achieved better performance compared to other projectors. These results demonstrate the effectiveness of the partial-global projection of our ParGo, which enhances the perception of image details while maintaining global perception. This dual capability leads to superior performance across diverse tasks.

Case Study

In Figure 3, we give several examples from the MM-Vet benchmark to demonstrate the superiority of our proposed Partial-Global projector in preserving partial and global information. Specifically, we compare our full model with the baseline that uses a Q-Former with 304 tokens as the projector. From the first two examples (in the first column), we



Q: What is the color of the hat worn by the person in the front left?

Q-Former : The person in the front left is wearing a **white** hat.

ParGo : The person in the front left is wearing a **blue** hat.

Q: What is the green logo on the car?

Q-Former : The green logo on the car is the **Ford logo**.

ParGo : The green logo on the car is the **Monster Energy logo**.

Q: What are all the scene text in the image?

Q-Former: The scene text in the image reads "Luxmi Jewellery."

ParGo : The scene text in the image is "Luxmi Jewellers."

Multi-Step Portrait: Temporal Directions

1. Before you draw ears, draw long curly hair.
2. Draw pink lips after you draw brown eyes.
3. Color the hair brown before you draw eyelashes.
4. Draw a nose after you draw earrings.
5. Before you draw a silver necklace, color the earrings green.

Q: What should be drawn in the fourth step?

Q-Former : In the fourth step of the multi-step drawing instructions, a pair of **eyes** should be drawn.

ParGo : In the fourth step, a **nose** should be drawn.

Q: Are all of the cats the same color?

Q-Former : Yes, all of the cats in the image are the same color, which is a mix of **orange, gray, and white.**

ParGo : **No**, the cats in the image are not all the same color. They are a mix of different colors, including **orange, gray, and brown.**

Q: What is pictured?

Q-Former : The place pictured is a beach.

ParGo : The image shows a beach **with a large elephant walking along the shoreline.**

Figure 3: Case study on the proposed Partial-Global projector (ParGo). In this figure, we select 6 examples to illustrate the superiority of our proposed ParGo in aligning vision and language modalities.

find that our model has a better perception of the location-specified partial object (*i.e.*, the person in the front left, the green logo on the car.), and correctly answers the question. Conversely, due to the Q-Former projector's overemphasis on prominent regions and neglect of image details, the baseline fails to answer the question correctly, *e.g.*, the blue Ford logo in the car. Regarding the two optical character recognition examples in the middle column, our model accurately identifies the characters in the image and correctly understands the text. In contrast, the baseline model fails to achieve correct recognition in these instances. The two examples in the last column illustrate that our method has a better perception of the overall images while maintaining the partial information, such as the colors of different cats and the presence of the elephant on the shoreline. In summary, the qualitative results highlight the ability of our proposed Projector to align the two modalities, providing the LLM with features covering both partial and global information.

Conclusion

In this work, we focus on the vision-language projector in MLLMs, proposing Partial-Global projector (ParGo). ParGo employs partial and global tokens with specially designed attention masks to extract two kinds of information separately, with considering the relation between different partial regions in an image. Moreover, to further facilitate the alignment between the two modalities, we contribute a large-scale detail-captioned dataset ParGoCap-1M-PT for pre-training. Extensive ablations and experiments are conducted, which illustrate the effectiveness of our ParGo. For instance, our ParGo outperforms Q-Former by 259.96 scores in the MME benchmark. We find that ParGo significantly outperforms other projectors, particularly in tasks that emphasize detail perception. These results highlight ParGo's potential to enhance MLLMs by providing a more nuanced understanding of visual content through the integration of both partial and global views.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2023. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *CVPR*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023a. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023b. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

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

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Dao, T. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. *arXiv preprint arXiv:2401.16420*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; and Ji, R. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; and Tu, Z. 2023. BLIVA: A Simple Multimodal LLM for Better Handling of Text-Rich Visual Questions.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*.

Liu, F.; Emerson, G.; and Collier, N. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Liu, Y.; Yang, B.; Liu, Q.; Li, Z.; Ma, Z.; Zhang, S.; and Bai, X. 2024b. TextMonkey: An OCR-Free Large Multimodal Model for Understanding Document. *arXiv preprint arXiv:2403.04473*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

OpenAI. 2023a. ChatGPT. https://chat.openai.com/.

OpenAI. 2023b. GPT-4 Technical Report.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rajbhandari, S.; Rasley, J.; Ruwase, O.; and He, Y. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.

Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; and Mottaghi, R. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of ACL*.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. 2023a. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023b. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. arXiv:2311.04257.

Yu, Q.; Sun, Q.; Zhang, X.; Cui, Y.; Zhang, F.; Cao, Y.; Wang, X.; and Liu, J. 2024a. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14022–14032.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024b. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.