
LMEye: An Interactive Perception Network for Large Language Models

Yunxin Li¹, Baotian Hu¹, Xinyu Chen¹, Lin Ma², and Min Zhang¹

¹Harbin Institute of Technology, Shenzhen

²Meituan, Beijing

liyunxin987@163.com, {hubaotian, zhangmin2021}@hit.edu.cn

forest.linma@gmail.com

<https://github.com/YunxinLi/LingCloud>

Abstract

Training a Large Visual Language Model (LVLM) from scratch, like GPT-4, is resource-intensive. Our paper presents a play-and-plug module for Large Language Models (LLMs), namely Interactive Perception Network (IPN)¹, aiming to achieve a LVLM by incorporating the image understanding capability into LLMs. Previous methods incorporate visual information into LLMs with a simple visual mapping network, where the image feature is projected into the embedding space of LLMs via a linear layer. Such mapping network projects the image feature once yet does not consider the interaction between the image and the human input query. Hence, the obtained visual information with no connections with human intention may be inadequate for LLMs to make intention-following responses, which we term as static visual information. IPN addresses this issue by allowing the LLM to request the desired visual information aligned with various human instructions, which we term as the dynamic interaction between the LLM and visual information. Specifically, IPN consists of a simple visual mapping network to provide the basic perception of an image for LLMs. It also contains additional modules responsible for acquiring requests from LLMs, performing request-based visual information interaction, and transmitting the resulting interacted visual information to LLMs, respectively. In this way, LLMs act to understand the human query, deliver the corresponding request to the request-based visual information interaction module, and generate the response based on the interleaved multimodal information. We evaluate IPN through extensive experiments on multimodal question answering, reasoning, and so on, demonstrating that it significantly improves the zero-shot performance of LVLMs on various multimodal tasks compared to previous methods.

1 Introduction

Vision-Language Models (VLMs) [1, 48, 49] trained on a massive amount of image-text data have shown impressive results in various multimodal understanding and generation tasks. However, training a LVLM (e.g., Flamingo [1], Kosmos-1 [18], and GPT-4 [36]) from scratch is resource-intensive. To alleviate this issue, previous open-source efforts [32, 23, 10, 16] present that we can construct a LVLM based on the text-only large language model (LLM) through transforming the visual information (obtained by frozen pretrained visual encoders [41, 9]) into the representation space of LLM. By doing so, LLMs are capable of understanding visual information and performing multimodal human-machine interaction. Significantly, the whole training process is parameter efficient since it

¹IPN is the main contribution of LMEye, i.e., incorporating the human-like eyes for LLMs.

only needs to optimize a few parameters of the vision-to-language feature transformer, similar to popular prefix or prompt tuning approaches [25, 19].

Recent work [32] demonstrates that a learnable linear mapping network can allow LLMs to incorporate the basic global perception information of an image. Different from common VLMs, e.g., Oscar [24] and OFA [48], LVLMs constructed in this way usually perform multimodal generation well [32] because LLMs are capable of powerful contextual understanding, reasoning, and generating capabilities. To step forward this direction, Koh et al. [20] present the model FROMAGe, where they freeze the LLM and visual encoder and fine-tune several linear mapping layers to achieve cross-modality information interactions. It realizes strong zero-shot performances on the contextual image retrieval and multimodal dialogue tasks. Li et al. [23] propose BLIP-2 with a lightweight Querying Transformer to bridge the vision and language semantic gap for frozen image encoders and large language models. In addition, the multimodal instruction-following tuning method is recently introduced by Liu et al. [27] and Zhu et al. [58] to advance the multimodal interaction capability of LLMs, which show supervisor performances on the multimodal dialogue scene.

However, for previous methods such as BLIP-2 and FROMAGe, the visual feature fed into the LLMs is transformed once via the visual mapping network and does not interact with human input queries, which we term as static visual information. Hence, the language model may not obtain the adequate visual information for the various queries. To address this issue, we present an interactive perception network (IPN) for LLMs, which allows the LLM to request the desired visual information aligned with various human instructions. Notably, we do not tune and modify the structure and parameters of LLMs. Otherwise, it will degenerate their original performances [10] on NLP tasks, thereby weakening the generalization of LLMs. The aim of this work is to enhance the multimodal information processing capability of LLMs without compromising their powerful language processing capability. In summary, the contributions of our proposed IPN lie in the following three-fold.

- We propose a novel IPN to make LLMs incorporate the desired visual information for various human queries. Specifically, IPN mainly consists of two modules: 1) the first one provides the basic global information of an image for LLMs, namely feature alignment; 2) the second one is responsible for acquiring the request from LLMs, performing request-based visual information interaction, and transmitting the resulting interacted visual information to LLMs.
- The whole training process of IPN is performed in two stages. Concretely, the first multimodal pretraining stage allows the LLM to obtain the basic global information of an image; the second multimodal instruction-following tuning stage mainly makes the overall information interactive workflow effective and adapt to various human queries.
- A benchmark is constructed to evaluate the performance of models on in-detail image description and VQA with long answer. Experimental results show that the proposed method can significantly improve the overall zero-shot performance on downstream tasks for various scales and types of LVLMs, e.g., exact gain by 5.01% on OK-VQA for IPN (BLIP-2), and exact gain by 20% on VQA with long answer for IPN (LLaMA-7b) vs. LLaVA (Vicuna-7b).

2 Related Work

Vision-assisted LLMs. Different vision-language models [1, 49, 18] which are trained from scratch with large-scale image-text pairs, vision-assisted LLMs is based on a pre-trained large language model, allowing it to understand visual information and be able to process multimodal information. They usually apply the recently proposed prefix-tuning [19, 25] or adapter-based [56] tuning methods to fine-tune the language model on specific multimodal tasks, so that they can be competent for some multimodal scenarios. For instance, Zhu et al. [59] utilized the text-to-image technical to generate the image and infused the visual information into the language model for multimodal text generation. Jing et al. [20] explored using the LLMs for image-text retrieval and multimodal text-image interaction. To step forward this direction, BLIP-2 [23] employs a Flan-T5 [8] or OPT [57] with a Q-Former to efficiently align visual features with the language model. Most recently, PaLM-E [10], featuring 562 billion parameters, has been developed to integrate real-world continuous sensor modalities into an LLM, thereby establishing a connection between real-world perceptions and human languages. To conclude, previous works demonstrate that it is a potential research direction for enabling frozen LLMs to handle multimodal information.

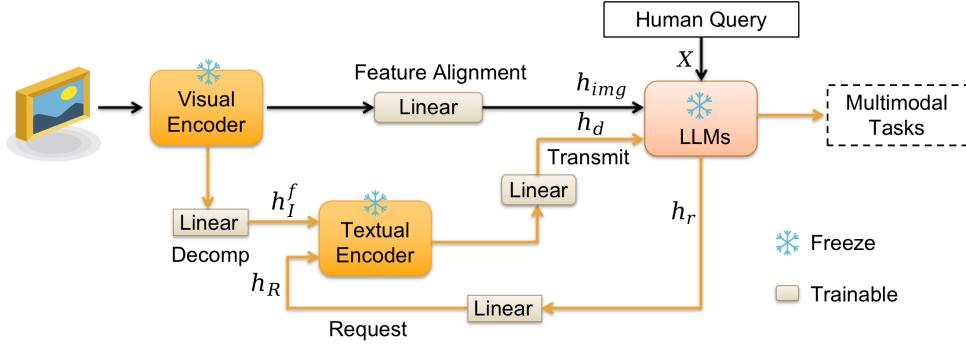


Figure 1: Illustration of the overall architecture of the Interactive Perception Network. Both Visual Encoder and Textual Encoder are from CLIP. The black lines show the first input to LLMs and orange lines represent the interaction stage between LLMs and visual information.

Multimodal Instruction-following Tuning. Advances in instruction-tuning text-only LLMs [37, 34, 8] have achieved impressive performance on the NLP tasks and human-machine interaction scenarios, such as Flan-T5, Bloomz [34], and ChatGPT. Recently, some researchers have explored using multimodal instruction data to fine-tune pre-trained LLMs to improve their multimodal human-machine interaction capability. Liu et al. [27] employs the GPT-4 to produce the multimodal instruction data and fine-tune the language model LLaMA [45] on the synthetic multimodal instruction-following dataset. Zhu et al. [58] also construct a well-aligned multimodal instruction-following dataset to fine-tune a robust instruction-tuned language model (Vicuna), and it achieves superior performance on open-domain multimodal dialogue. Zhang et al. [56] present a lightweight adaption method to efficiently fine-tune LLaMA into an instruction-following model. In this paper, we introduce various multimodal instruction data to make the IPN adapt to open-domain multimodal scenarios.

Vision Tools for LLMs. A recent line of research [33, 40] investigates ways to enhance the performance of LLMs by enabling them to access external tools such as vision foundation models, search engines, or other APIs for solving complex problems. This approach broadens the scope of LLMs in processing information of varying complexities. For example, Toolformer [42] empowers LLMs to decide which APIs to use, when to use them, what arguments to pass, and how to incorporate the resulting information into text generation. Low-code LLM [3] uses six simple low-code visual programming interactions, such as clicking, dragging, or text editing, to achieve more controlled and reliable responses. In contrast, Lu et al.[30] propose a plug-and-play compositional reasoning framework Chameleon that augments LLMs to address complex challenges, such as using off-the-shelf vision models. Wu et al.[52] introduce visual Chatgpt, which designs a set of prompts to incorporate visual model information into ChatGPT, considering models with multiple inputs/outputs and those that require visual feedback. Unlike the above pipeline approaches, our work focuses on end-to-end interaction framework between LLMs and visual information.

3 Interactive Perception Network

3.1 Architecture

As shown in Figure 1, the architecture of the interactive perception network contains four learnable linear layers, which are response for different functions. Given an image I and one human’s query $X = (x_1, \dots, x_M)$, where x_i represents the i -th token of the text query input to LLMs, we obtain the global image feature $h_I \in \mathbb{R}^{1 \times d_v}$ by the pretrained visual encoder of CLIP-ViT-L/14 [41]. d_v is the hidden state size of visual encoder. Meanwhile, a learnable special token $\langle img \rangle$ is added in the word embedding table of LLMs as the input position tag of image feature, which will be used in the following feature alignment process.

Feature Alignment. We employ a linear projection layer as the visual mapping network to project the global image feature into the language embedding space. Such a process $f(h_I) = W \cdot h_I + b$ is named as the feature alignment, where $W \in \mathbb{R}^{d_v \times d_L}$ and b are learnable parameters (similar to following

linear layers) and d_L is the hidden state size of LLMs. By doing so, LLMs can obtain the basic/global perception information of an image, which is added with the presentation of token $\langle img \rangle$. We also add another special token $\langle img-d \rangle$ at the end of the image and human query sequence to capture the whole encoding information of image and human query. Hence, the first input sequence of LLMs can be denoted to $(\langle img \rangle, f(h_I), X, \langle img-d \rangle) \rightarrow (h_{img}, h_X, h_{img-d})$, where h_{img} refers to the addition of $f(h_I)$ and the token representation of $\langle img \rangle$. h_X and h_{img-d} are the corresponding word encoding representations of X and $\langle img-d \rangle$. The final output of the token $\langle img-d \rangle$ at the last layer of LLMs is expected to contain the semantic meaning of human query, which is present to $h_r \in \mathbb{R}^{1 \times d_L}$, since previous works [35, 36, 45] have shown that recent LLMs have been able to understand various human languages. In addition, h_r may also contain the image content via the self-attention mechanism of LLMs, yet we argue that text-only LLMs without pretraining on multimodal data cannot incorporate visual information [32] well like powerful pretrained multimodal image and language models. To facilitate LLMs incorporating the desired visual information aligned with a human query, instead of optimizing the parameters of LLMs (with full-parameter or the Low-rank adaptation [17]) on specific data, like LLaVA [27] and mPLUG-Owl [53], we perform the interaction between human query and visual information outside the LLMs. In this way, LLMs could still maintain their original powers and generalization on natural language tasks because the structure and parameters of the language model are not changed.

Request-based Visual Information Interaction. First, we apply a linear project layer to map the above hidden state h_r into the space of the following information interactive module, denoted to h_R . We regard this process is acquiring the request from LLMs, e.g., the LLMs encoding request information of human query “*What is the man doing in the image?*”. After gaining the request of LLMs, we propose utilizing h_R and image features to perform multimodal information interaction. To this end, we adopt the pretrained textual encoder of CLIP-ViT/14 as the backbone of information interaction, since the textual encoder has the similar embedding space with the image encoder. As the bottom part shown in Figure 1, we first decompose the original global encoding of an image into fine-grained ones by a linear layer. The length of fine-grained image features is set to 5, where we have no special consideration and employing the fine-grained image patch features may have better performance yet with greater computational overhead. Then, we concatenate fine-grained image features $h_I^f = (h_I^{f_1}, \dots, h_I^{f_5})$ and requested information h_R into the frozen textual encoder of CLIP to perform interaction. Specifically, the input sequence is $\langle h_I^f, h_R, h_g \rangle$, where h_g is a global vector to capture the whole multimodal interaction information and is initialized by the word vector of [SEP] in the embedding table of the textual encoder. The output of the last Transformer layer for h_g is regarded as the resulting interacted visual information and will be transmitted to the LLM via a linear layer. As shown in Figure 1, the final output is denoted to h_d . Afterwards, the new presentation sequence (h_{img}, h_X, h_d) is fed into the LLM to generate the final response.

3.2 Training

The whole training process includes the following multimodal pretraining and instruction-following tuning stages. The LLM, visual encoder, and textual encoder are frozen in the whole training process.

Phase 1: Multimodal Pretraining. This stage aims to train the linear project layer in feature alignment. We adopt tens of millions of released image-text pairs from CC3M [4], LAION-400M [43], COCO Caption [6], and Flickr3K [39]. The total number of image-text pairs is about 69 Million because some images from CC3M can not be downloaded and partial images from the LAION-400M are used. Notably, introducing more high-quality image-text data will further improve the effectiveness of feature alignment, yet it is not the current focus of this work. Suppose that the caption (training target) of an image is $Y = (y_1, \dots, y_N)$, where y_i represents the i -th token and N refers to the total length, the optimizing loss of this stage is as following:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log P_i(\hat{y}_i = y_i | h_{img}; y_1, \dots, y_{i-1}). \quad (1)$$

Phase 2: Multimodal Instruction-following Tuning. This stage mainly makes IPN effective by using various multimodal instruction-following data. First, we construct two types of image-text semantic matching data based on the image-text pairs from datasets CC3M, COCO Caption, and Flickr3k. As shown in Figure 2, the two types are “True or False” inference and four-choice selection

Multimodal Instruction-following Data	
	<p>Type1: Image: \n Caption: Closeup of bins of food that include broccoli and bread.\nQuestion: Is this caption matching the image ?\nOptions: True, False <img-d>\nAnswer: True</p> <p>Type1: Image: \nCaption: A beautiful kitchen and counters with lit lights.\nQuestion: Is this caption matching the image?\nOptions: True, False <img-d>\nAnswer: False</p> <p>Type2: Image: \nQuestion: Which of the following caption is correct?\nOptions: A: Three people smiling for the camera beside a bike. B: Closeup of bins of food that include broccoli and bread. C: People walking and sitting along a road dividing a green park and a cemetery. D: A bike parked next to a bench with two people on it. <img-d>\nAnswer: B</p>
	<p>Image: \nQuestion: What is the primary action occurring in the image? <img-d>\nAnswer: The primary action occurring in the image is a baseball player swinging at a pitched ball during a baseball game.</p> <p>Image: \nQuestion: What is the role of the man with the baseball bat? <img-d>\nAnswer: The man with the baseball bat is a hitter trying to make contact with the baseball and score runs for his team.</p> <p>Image: \nQuestion: What is the position of the baseball player on the field? <img-d>\nAnswer: Answer: The baseball player is at home plate, which is where the hitter stands to swing at the ball.</p>
	<p>Image: \nHuman: Write a detailed description of the given image. <img-d>\nAnswer: The image displays a young man and woman standing in a grassy area in what appears to be a city park. Both are drinking bottled water, with the woman holding her bottle close to her, and the man holding his above his waist. A frisbee is visible on the ground nearby, suggesting they may have been engaged in some outdoor activities. Cars are parked in the background, and an airplane can be seen flying in the sky above the grassy area. There are buildings and trees visible in the distance, further emphasizing the park-like setting.</p>

Figure 2: Visualization of the multimodal instruction-following tuning data. and <img-d> are special tokens used to incorporate static visual information of an image and drive request-based visual information interaction, respectively.

tasks, respectively, where captions are randomly sampled from the corresponding training set. By doing so, the overall IPN could be trained to help and improve LLMs performing image-text semantic alignment. Second, to make IPN adapt to various human queries, we introduce the multimodal instruction-following data about conversation and complex reasoning, released by Liu et al. [27]. Finally, considering that a complex image contains infinite-level visual information and may be attached to external knowledge, as the last example shown in Figure 2, we introduce data about the in-detail description of an image to improve the multimodal long text generation capability of IPN. It includes the corresponding data from Liu et al. [27] and the artwork description dataset SemArt [13]. The total number of all instruction data is about 7.3M, encompassing 7.1M semantic matching data, 20k artwork analysis samples, and 150k additional samples.

4 Experiment

In this section, we will evaluate the designed IPN on the following multimodal scenarios: Visual Question Answering, Multimodal Reasoning, and In-detail Image Description.

4.1 Experimental Settings

Datasets. First, we will evaluate IPN and other visual language models on three visual understanding and reasoning datasets: validation sets of VCR [54] and VQAv2 [15], and the test set of OK-VQA [31]. We also evaluate our model LMEye (IPN) on the recently released multimodal evaluation benchmark MMBench [28]², which is a systematically-designed objective benchmark for robustly evaluating the various abilities of vision-language models. In addition, we also use the GPT-3.5-turbo [35] to produce five question-answer pairs centered around an image based on about 3.5k images and their long descriptions from Zhu et al. [58]. The prompt template is “*Generate five question-answer pairs for the following image description. Demand: the answer of question must be contained in the description and the type is Question: ... Answer: ... \n Description:* ”. The total number of question-answer pairs is about 17.5k wherein the length of the answers exceeds that of conventional VQA datasets, with an average length of 13 words, like the second image and its question-answer pairs depicted in Figure 2. The constructed data will be used to evaluate and analyze the performance of models.

²<https://opencompass.org.cn/leaderboard-multimodal>

Table 1: Model performances on MMBench. “TotalParams” indicates the total parameters of models. Logical Reasoning (LR), Attribute Reasoning (AR), Relation Reasoning (RR), Fine-grained Perception (Cross Instance) (FP-C), Fine-grained Perception (Single Instance) (FP-S), and Coarse Perception (CP)

Models↓ Types →	TotalParams	Overall	LR	AR	RR	FP-S	FP-C	CP
OpenFlamingo [1]	9B	4.3	9.1	11.4	3.3	2.5	1.6	1.5
OpenFlamingo v2 [1]	9B	5.7	11.4	12.8	1.4	5.5	0.8	4.0
MiniGPT-4 [58]	8B	23.0	13.6	32.9	8.9	28.8	11.2	28.3
MMGPT [14]	9B	16.0	1.1	23.8	20.7	18.3	5.2	18.3
PandaGPT [44]	14B	30.6	15.3	41.5	22.0	20.3	20.4	47.9
VisualGLM [11]	8B	33.5	11.4	48.8	27.7	35.8	17.6	41.5
InstructBLIP [27]	8B	33.9	21.6	47.4	22.5	33.0	24.4	41.1
LLaVA [27]	7.2B	36.2	15.9	53.6	28.6	41.8	20.0	40.4
LLaMA-Adapter-v2 [12]	7.2B	38.9	7.4	45.3	19.2	45.0	32.0	54.0
G2PT	7B	39.8	14.8	46.7	31.5	41.8	34.4	49.8
mPLUG-Owl [53]	7.2B	56.7	30.7	65.7	52.1	61.0	45.6	65.1
Otter-I [22]	9B	48.3	22.2	63.3	39.4	46.8	36.4	60.6
Shikra [5]	7.2B	60.2	33.5	69.6	53.1	61.8	50.4	71.7
IPN (ours)	4.4B	61.3	36.9	73.0	55.4	60.0	58.0	68.9

Comparing Models. Flamingo [1] is a unifying multimodal generative model capable of rapidly adapting to a variety of image and video tasks. OFA [48] is a sequence-to-sequence learning framework that could unify a diverse set of cross-modal and unimodal tasks. FROMAGe [20] is a typical LVLM that is efficiently trained by visually grounding LLMs with image captioning and contrastive learning, capable of image caption and image-text retrieval. BLIP-2 [23] is a two-stage training strategy that bootstraps vision-language representation learning and vision-to-language generative learning based on the frozen image encoder and language model, achieving state-of-the-art performances on various multimodal tasks. In addition, we also compare our methods with multimodal instruction-tuned model MiniGPT-4 [58] and LLaVA [27], where MiniGPT-4 is based on the pretrained Q-former from BLIP-2. Compared to BLIP-2 and FROMAGe, they are tuned with the multimodal instruction-following data generated by GPT-4. During the multimodal instruction tuning stage, both the projection matrix and LLM of LLaVA are updated.

Implementation Details. We run all experiments on eight Tesla A100-80G GPUs with the Python environment. To verify the effectiveness of the proposed IPN, we adopt OPT-iml-1.3b, Bloomz-7b1, LLaMA-7b/13b [45], and BLIP-2 (FlanT5_{XL}) as the backbone of our framework. For the multimodal pretraining stage, we set the initial learning rate to 1e-4 and use the AdamW [29] optimizer to optimize the feature alignment process with the cosine declining way. The total training step of this stage is one epoch, and the batch size is 768. During the multimodal instruction tuning stage, we adopt a smaller batch size (256) and set the initial learning rate to 1e-4. We will freeze the first-stage parameters (include the linear project layer in feature alignment and the token representation of **, or Q-former in BLIP-2) while performing multimodal instruction tuning. During generation, we employ the beam sample generation strategy from HuggingFace Transformer³ repository and set the beam sizes to 5 and 1 for in-detail image description generation and VQA, respectively.

Evaluation Metrics. For visual question answering (VQA) with short answer and visual reasoning datasets, we adopt the common EM (exactly matching) calculation way as the evaluation method of accuracy. For the in-detail image description generation and VQA with long answer, we employ several generative evaluation metrics: BLEU [38], ROUGE [26], CIDEr [46], and METEOR [2].

³<https://github.com/huggingface/transformers>

Table 2: Model performances on SEED-Bench [21]. We evaluate IPN (FlanT5-XL)-4.4B on 9 dimensions for image understanding, including Scene Understanding (SU), Instance Identity (II), Instance Location (IL), Instance Attribute (IA), Instance Counting (IC), Spatial Relation (SR), Instance Interaction (IIR), Visual Reasoning (VR), and Text Recognition (TR). Baseline results are from SEED-Bench.

Models↓ Types →	Overall	SU	II	IL	IA	IC	SR	IIR	VR	TR
OpenFlamingo v2 [1]	34.51	43.86	38.12	31.28	30.06	27.30	30.59	29.90	50.15	20.00
MiniGPT-4 [58]	47.40	56.27	49.15	45.82	37.93	45.32	32.57	47.42	57.10	11.76
MMGPT [14]	34.54	43.64	37.85	31.45	30.78	27.34	30.14	29.90	51.36	18.82
BLIP-2 FlanT5-XL [23]	49.74	59.12	53.90	49.19	42.33	43.15	36.68	55.67	45.62	25.88
InstructBLIP FlanT5-XL [27]	57.80	60.29	58.49	63.37	40.59	58.44	38.66	51.55	45.92	25.88
InstructBLIP Vicuna [27]	58.76	60.20	58.93	65.63	43.56	57.05	40.33	52.58	47.73	43.53
LLaVA [27]	36.96	42.69	34.90	33.45	28.43	41.85	30.75	27.84	46.83	37.65
LLaMA-Adapter-v2 [12]	35.19	45.22	38.50	29.30	33.03	29.67	35.46	39.18	51.96	24.71
GVT [47]	35.49	41.74	35.50	31.79	29.45	36.17	31.96	31.96	51.06	27.06
VPGTrans [55]	41.81	51.87	44.13	39.90	36.09	33.71	36.38	31.96	53.17	30.59
mPLUG-Owl [53]	37.88	49.68	45.33	32.52	36.71	27.26	32.72	44.33	54.68	18.82
Otter [22]	35.16	44.90	38.56	32.24	30.88	26.28	31.81	31.96	51.36	31.76
IPN (ours)	59.70	73.20	64.12	56.57	53.99	48.75	47.64	65.98	76.13	37.65

Table 3: Zero-shot performances on some common multimodal datasets. IPN variants with “*” indicate that we only remain the first linear layer in feature alignment and remove the interactive process. “NumImg” represents the total number of images contained in the multimodal pretraining stage.

Models↓ Types →	NumImg	VCR (Q→A)	VQAv2	OK-VQA
Flamingo-3B [1]	>1B	-	49.2	41.2
MiniGPT-4 (Vicuna-7b) [58]	129M	-	44.31	32.16
LLaVA (Vicuna-7b) [27]	0.6M	-	56.59	50.42
OFA-Large [48]	20M	25.07	40.25	19.34
FROMAGe (OPT-6.7b) [20]	3.3M	20.87	44.08	20.06
BLIP-2 (FlanT5 _{XL}) [23]	129M	57.30	59.95	48.91
IPN (OPT-iml-1.3b)*	1.7M	34.34	38.34	22.26
IPN (OPT-iml-1.3b)	1.7M	39.52	42.42	24.58
IPN (Bloomz-7b1)*	13M	39.31	40.63	25.56
IPN (Bloomz-7b1)	13M	43.07	42.20	26.38
IPN (Bloomz-7b1)*	69M	42.81	42.39	26.79
IPN (Bloomz-7b1)	69M	47.00	45.58	35.11
IPN (BLIP-2, FlanT5 _{XL})	129M	57.40	57.60	53.96

4.2 Results and Analysis

MMBench Evaluation Results. In addition, we have augmented the instruction data for IPN by introducing a training set of approximately 20 multimodal tasks. This is aimed at improving the generalization of IPN. The evaluation results on the MMBench are presented in Table 1. The results show that our proposed model, LMEye, outperformed other models on the overall evaluation leaderboard while using fewer parameters. It is worth noting that LMEye outperformed other models in terms of reasoning performance, particularly in Logical Reasoning (LR), Attribute Reasoning (AR), and Relation Reasoning. This indicates that LMEye (IPN) is capable of effectively reasoning and making connections between different pieces of information, leading to superior performance compared to other models.

Table 4: Model performances on the self-constructed evaluation benchmark.

Models↓ Types →	NumImg	VQA with Long Answer					
		B@1	B@2	R@1	R@L	CIDEr	METEOR
LLaVA (Vicuna-7b)	0.6M	29.23	17.62	40.85	38.67	166.95	50.15
IPN (Bloomz-7b1)*	13M	2.15	0.40	7.32	7.24	7.49	4.47
IPN (Bloomz-7b1)	13M	37.49	23.31	34.94	33.41	131.83	46.61
IPN (Bloomz-7b1)*	69M	3.08	0.52	7.08	7.0	7.51	4.67
IPN (Bloomz-7b1)	69M	34.77	22.0	39.20	37.37	167.17	48.60
IPN (LLaMA-7b)	1.7M	42.40	27.84	41.06	39.37	188.59	50.22
IPN (LLaMA-13b)	1.7M	49.36	33.67	43.96	42.05	212.16	51.24
BLIP-2 (FlanT5XL)	129M	10.60	5.43	25.29	24.21	61.88	17.75
IPN (BLIP-2, FlanT5XL)	129M	41.56	26.49	47.14	44.46	193.92	50.63
In-detail Image Description							
Models↓ Types →	NumImg	B@1	B@2	R@1	R@L	CIDEr	METEOR
IPN (Bloomz-7b1)*	13M	10.55	1.73	5.78	5.60	0.8	7.02
IPN (Bloomz-7b1)	13M	27.50	6.0	12.0	11.68	2.4	21.15
IPN (Bloomz-7b1)*	69M	11.59	2.37	6.92	5.68	0.6	11.15
IPN (Bloomz-7b1)	69M	26.27	6.71	14.15	13.80	2.5	26.71
IPN (LLaMA-7b)	1.7M	26.21	6.47	11.18	10.86	1.4	26.02
IPN (LLaMA-13b)	1.7M	29.91	7.11	12.06	11.79	3.8	23.06
BLIP-2 (FlanT5XL)	129M	1.21	0.56	18.37	17.48	2.3	8.15
IPN (BLIP-2, FlanT5XL)	129M	8.98	2.47	17.06	16.17	3.0	14.05

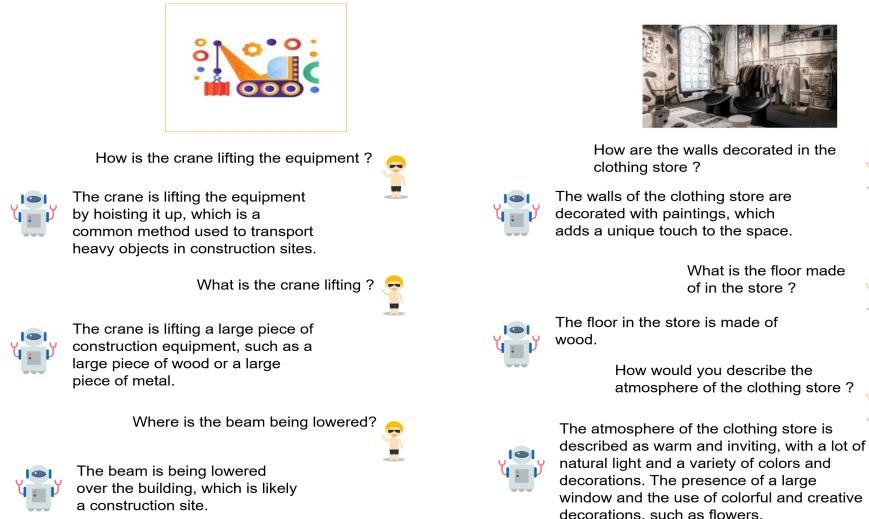


Figure 3: Illustration of images with multiple questions and answers generated by IPN (Bloomz-7b1).

SEED-Bench Evaluation Results. The experimental results presented in Table 2 demonstrate the effectiveness of LMEye in achieving state-of-the-art (SOTA) performance. Specifically, LMEye has shown significant improvements in scene understanding, with an increase of 13 points compared to the previous SOTA. Moreover, in the category of sample attribute recognition and spatial connection understanding, LMEye has outperformed InstructBLIP. These results highlight the effectiveness of a plug-and-play interactive perception framework IPN, in enhancing the ability of language models to understand images. Overall, these findings demonstrate the potential of language models in advancing the field of image understanding and suggest that plug-and-play interactive perception frameworks like the IPN can be an effective means of leveraging these capabilities. Further research in this area may pave the way for more sophisticated and effective approaches to image understanding, with implications for a wide range of applications and industries.

Images:				
Models				
IPN (LLaMA-7b)	Q: What color is the man's shirt? The man is wearing a white shirt.	Q: What kind of birds are these? These are pelicans.	Q: What color is the wall? The wall is white.	Q: What kind of animal is on the yellow flag? A dog. ✗
IPN (LLaMA-13b)	The man's shirt is white.	The birds are seagulls.	The wall is white.	The animal on the yellow flag is a lion.
IPN (Bloomz-7b)	The man's shirt is white.	These birds are seagulls.	The wall is white.	The yellow flag in the background is a lion.
BLIP-2 (FlanT5xl)	Blue ✗	Birds ✗	Purple ✗	Elephant ✗
LLaVA (Vicuna-7b)	The man's shirt is red. ✗	The birds are a flock of flamingos. ✗	The wall is purple. ✗	There is a horse on the yellow flag. ✗
MinIGPT-4 (Vicuna-7b)	The man's shirt is blue. ✗	These birds appear to be flamingos, a type of water bird with pinkish-red feathers. ✗	The wall appears to be made of bricks and is grey in color.	The image shows a group of cyclists on a road race, with a large crowd of spectators watching from the side of the road. ✗

Figure 4: Several results generated by IPN variants and some comparative models. The response with the symbol “✗” are incorrect.

In-deail Description Gneneration	
	<p>Author: DYCK, Sir Anthony van Title: Portrait of a Man in Armour with Red Scarf Technique: Oil on canvas Date: 1625-27 Type: portrait School: Flemish Prompt: Write a detailed description of this image</p> <p>We are likely to have anthony van dyck as a landscape painter of standing in the first decade of the seventeenth century. his allegorical works are usually more modest. he had received classical instruction but his own conception was original and vigorous, even though he had mastered a loose, local, tradition. in a series of biographical portraits, between 1628 and 1629, van dyck exploited the possibilities of the early flemish portraiture. this was also his first attempt at portraiture with a dramatic narrative presentation; he took up the problem of the golden figure already early in his career and this particular image clearly shows how far he had already come. the face and armour of the nude form a single unit; a red scarf (with the words 'peasants' on it) contains the reference to the 'ride of the schips', a drama recorded in the journals of the time.</p>
	<p>Prompt: Give a detailed description of this image.</p> <p>The photo depicts a crowded, traffic-filled street scene with multiple people crossing the street in Washington, D.C. The people walk in different directions, varying from north to south, west to east, and in front of or behind others. It can be assumed that the street scene is unfamiliar to them, as the people are unfamiliar with their surroundings. Beyond the street, a parked car can also be spotted in the scene, providing an interesting contrast to the extremely busy city streets. In the distance, tall buildings can also be seen surrounding the scene. This high-rise area may offer an alternative viewpoint for the people in the street, compared to the urban surroundings.</p>

Figure 5: Illustration of two cases about detailed image descriptions generated by IPN (Bloomz-7b1).

Visual Question Answering and Multimodal Reasoning. The experimental results are present in Table 3. In Table 3, we do not present the VCR result of LLaVA and MiniGPT-4 since they do not follow the prompt to select one option from four candidates. This may be attributed to the self-ability of Vicuna [7], which is the LLaMA tuned only with conversation data. Compared to conventional VLMs such as Flamingo-3B and OFA, the designed IPN variants and other LVLMs can be able to achieve better zero-shot performance on answer selection (VCR) and short answer generation tasks (VQA), even in the case that IPN (Bloomz-7b1) have only seen 1.7M images during the pretraining stage. It indicates that in the case of low resources, it is feasible to adopt such a training efficient LVL construction method based on the frozen visual and language models. In addition, introducing more powerful language models and high-quality image-text data in the pretraining stage will improve the accuracy of the language model in understanding image information, e.g., the performance comparisons of various IPN (Bloomz-7b1) and IPN (OPT-iml-1.3b) variants. When we introduce IPN for BLIP-2 and train the new framework through the collected multimodal instruction data, it substantially improves the performance on the complex visual problem task OK-VQA by about 5%. The main reason for the decline in performance on VQA may be attributed to the fact that the evaluation metric is word matching and IPN (BLIP-2) generates longer responses than BLIP-2, like the case shown in Figure 3. The ablation experiments of IPN with different LLMs and BLIP-2 (LVL) show that the proposed interactive perception network could be play-and-plug in various large language models and improve the performance of well-established BLIP-2 by incorporating request-based interacted visual information.

VQA with Long Answer. We mainly evaluate various IPN variants on the constructed evaluation benchmark: in-detail image description and visual question-answering tasks. The experimental results are shown in Table 4. According to the experimental results, we can see that multimodal instruction-tuned IPN models significantly improve almost all generative metrics. Combined with the given examples from this dataset shown in Figure 3, we suggest that the multimodal instruction-following tuning approach is helpful for LLMs to achieve an image understanding capability similar to GPT-4. From the cases shown in Figure 3, we find that IPN (Bloomz-7b1) could understand the intent of various questions and generate an accurate response. Compared with LLaVA (Vicuna-7b), MiniGPT-4, and BLIP-2, which incorporate the static visual information for different questions about an image, our method can obtain corresponding visual information relevant to the human input question (query) and generate more accurate responses (see the comparative performance shown in Table 4 and Figure 4). Moreover, the overall learnable parameters for IPN variants with four linear layers are about 25M, less than LLaVA and other models.

In-detail Image Description. Unlike the multimodal instruction-following data used by MiniGPT-4 and LLaVA, we introduce the artwork description data as part of instruction-tuning data and thus improve the model’s capability to understand the artwork. Judging from the sample shown in Figure 5, we observe that LLMs could utilize their stored knowledge to present sufficient analysis for the artwork. Compared to the artwork analysis, the in-detail description of a natural image focuses more on the comprehensiveness of visual information, e.g., the bottom description illustrated in Figure 4 mainly presents the salient aspects of image. From Table 4, we also observe that the improvement of detailed image description ability mainly comes from using relevant instruction data. Our method mainly improves the performance of LVLMs on VQA tasks with various queries. We present more cases in Appendix.

5 Discussion and Future Work

We mainly present some observations and future directions for multimodal instruction-tuning LVLMs:

Instruction-tuned LLMs can better generalize on multimodal tasks. Previous work [23] shows that the BLIP-2 variant with instruction-tuned LLMs performs best on many multimodal tasks. In Table 3, we observe that IPN (OPT-iml-1.3b)* is capable of better performance on VCR and OK-VQA compared to FROMAGE (OPT-6.7b) with a larger-scale OPT version. This can be attributed to the fact that text-only instruction-tuned LLMs better understand human input queries than original LLMs. Thus they have better performances on multimodal QA tasks.

Quality and Diversity of multimodal instruction-following data are important. A comparison between LLaVA (Vicuna-7b) and MiniGPT-4 (Vicuna-7b) reveals that LLaVA, which incorporates a larger number of diverse multimodal instruction data, outperforms MiniGPT-4. This finding is consistent with the research conducted by Liu et al. [27], who demonstrate that diverse multimodal instruction data can enhance the overall performance of LVLMs across various tasks. Current multimodal instruction-following data are usually constructed by powerful GPT-4 through the Self-Instruct [50] technical. While these automatically generated instruction data exhibit diversity, there remains room for improvement in terms of quality. In the future, it would be advantageous to incorporate high-quality multimodal task data, including video, image, and audio, to enhance the comprehensive capability of instruction-tuned LVLMs.

Visual information should interact with human instruction. Previous work BLIP-2 [23] attempts to input textual questions into the Q-former to refine its performance in the specific visual question-answering task, leading to superior results. These questions facilitate visual information extraction by utilizing self-attention layers within the Q-Former architecture. Different from BLIP-2, the proposed IPN focuses on extracting image features that are highly informative for the encoding request from LLMs, achieving dynamic interaction between LLMs and visual information. Additionally, we introduce diverse multimodal instruction-following data to train IPN-based LVLMs, allowing them to adapt to a wide range of human queries. Consequently, the LLMs can leverage enriched visual information to accomplish different tasks effectively. To conclude, enabling visual information to interact with human instruction/query is effective for improving the capability of LVLMs.

Hallucination. While LVLMs generate the in-detail image description or artwork analysis, they easily produce fragments that are nonsensical or unfaithful to the objective image and common knowledge, or fabrication of facts. To address this issue, in the future, we can introduce the alignment technical

(such as Reinforcement Learning from Human Feedback (RLHF) [37]), retrieval augmentation, or multimodal chain-of-the-thought (COT) [51] to improve the factuality of generated content.

6 Limitation

While our models strive to enhance their alignment with human queries, it is important to acknowledge that they are not completely aligned nor entirely safe. Despite efforts to improve the quality of outputs, our models still have limitations in avoiding generating toxic or biased content, fabrication of facts, and other undesirable outputs. In some cases, the model may inadvertently generate offensive, discriminatory, or harmful outputs. It can be attributed to biases in the training data or the self-ability of LLMs. Furthermore, due to constraints in the quality and diversity of available multimodal instruction-following data, the model may provide incorrect responses for certain queries.

7 Conclusion

We present IPN, an interactive perception network for large language models, aiming to achieve a large visual language model with the dynamic interaction between LLMs and visual information. The experimental results show that our method achieves superior zero-shot performances on visual question answering, in-detail image description, and multimodal reasoning tasks.

References

- [1] 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.
- [2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.
- [3] Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, Wang You, Ting Song, Yan Xia, et al. Low-code llm: Visual programming over llms. *arXiv preprint arXiv:2304.08103*, 2023.
- [4] 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.
- [5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.
- [6] 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.
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [11] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of*

- the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.
- [12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
 - [13] Noa Garcia and George Vogiatzis. How to read paintings: Semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference in Computer Vision Workshops*, 2018.
 - [14] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
 - [15] 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.
 - [16] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *CVPR*, 2023.
 - [17] 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. *arXiv preprint arXiv:2106.09685*, 2021.
 - [18] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
 - [19] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022.
 - [20] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.
 - [21] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
 - [22] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.
 - [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023.
 - [24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXXI*, pages 121–137. Springer, 2020.
 - [25] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*, 2021.
 - [26] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.
 - [27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
 - [28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
 - [29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
 - [30] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

- [31] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvpr conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- [32] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *ICLR*, 2023.
- [33] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.
- [34] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.
- [35] OpenAI. Chatgpt. *OpenAI Blog*, 2023.
- [36] OpenAI. Gpt-4 technical report. <https://arxiv.org/abs/2303.08774>, 2023.
- [37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [39] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.
- [40] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*, 2023.
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [42] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.
- [43] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [44] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.
- [45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [46] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.
- [47] Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*, 2023.
- [48] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.
- [49] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.
- [50] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

- [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [52] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [53] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [54] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.
- [55] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023.
- [56] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.
- [57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuhui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [58] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [59] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Visualize before you write: Imagination-guided open-ended text generation. *EACL*, 2022.

A Visual Question Answering

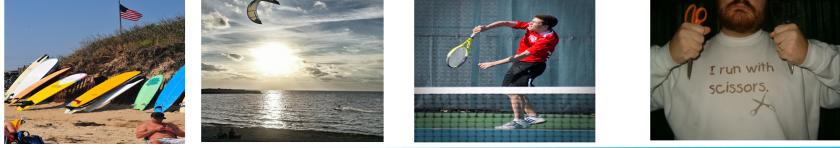
Images: 	Models IPN (LLaMA-7b) IPN (LLaMA-13b) IPN (Bloomz-7b) BLIP-2 (FlanT5xl) LLaVA (Vicuna-7b) MiniGPT-4 (Vicuna-7b)	Q: What color is the man's cap? The man's cap is black. The man's cap is black. The man's cap is black. Blue ✘ The cap is blue. ✘ The color of the cap is blue. ✘	Q: What sport is this? This is a surfing image This sport is paragliding. The sport is surfing. Sailing ✘ The image depicts a beach scene where people are enjoying various activities. ✘ This is a picture of a kiteboarder riding the waves in the ocean during sunset.	Q: Does his shoes have stripes or polka dots? Yes, his shoes have stripes. His shoes have stripes. His shoes have stripes. polka dots ✘ The young man's shoes have polka dots on them. ✘ The image is a man in a red shirt and white shorts holding a tennis racket ✘	Q: Does the man appear joyful or irritated? The man appears to be irritated in the image. The man appears irritated. The man appears irritated. joyful ✘ The man appears to be joyful, as he is smiling while holding a pair of scissors. ✘ The man in the image appears to be holding scissors and has a smile on his face, so he appears to be happy ✘
--	---	--	--	--	---

Figure 6: Illustration of VQA cases generated by several models.

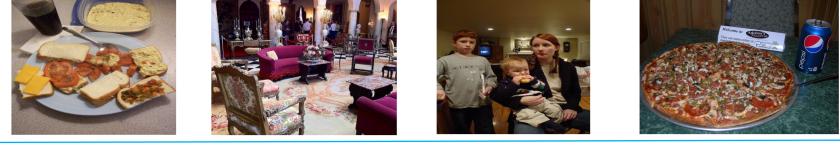
Images: 	Models IPN (LLaMA-7b) IPN (LLaMA-13b) IPN (Bloomz-7b) BLIP-2 (FlanT5xl) LLaVA (Vicuna-7b) MiniGPT-4 (Vicuna-7b)	Q: What is the color of the table? The table is white. The table is white. The table is white. Blue ✘ The table is blue. ✘ The color of the table is blue. ✘	Q: What color are the tall vases? The tall vases are white. The tall vases by the pillars are white. The tall vases by the pillars are white. Pink ✘ The tall vases are gold in color. ✘ The tall vases in the image are pink. ✘	Q: How old is the boy? The boy is a young child. The boy is a baby. The boy is a baby. three ✘ The boy is three years old. ✘ I'm sorry, I am unable to provide the image you requested, but I can give you some information	Q: Which utensil is there? There is a fork in the image. There is a fork in the image. There is a fork in the image. spoon There is a large pizza on a pan sitting on a table. ✘ A pizza ✘
--	---	---	---	--	---

Figure 7: Visualization of additional VQA cases generated by several models.

B Multi-turn Question Answering for One Image

 BABY Care	Question: What does the logo feature? The logo features a heart, which is a common symbol of love and affection.
	Question: How would you describe the design of the logo? The design of the logo is simple and clean, featuring a white background with a red heart and a black and white image of a woman holding a baby.
	Question: What type of company is the logo best suited for? The logo is best suited for a company that focuses on health and wellness, such as a gym, fitness center, or a health clinic. The logo features a woman's image, which can be associated with health and wellness, as well as the concept of beauty and fitness. The image of the woman can also be used to promote a sense of empowerment and confidence, which are important aspects of a healthy lifestyle.

Figure 8: A multi-turn question answering case generated by IPN (LLaMA-7b).



- 👤 Question: What does the image depict?
🤖 The image depicts a beautiful sunset scene with a mountain in the background.
The sun is setting behind the mountain, creating a stunning view.
👤 Question: How does the tree and birds in the background add to the image?
🤖 The tree and birds in the background add to the image by providing a sense of nature and a connection to the outdoors. The tree provides a sense of stability and a natural setting, while the birds add a sense of life and movement to the scene. The combination of these elements creates a peaceful and serene atmosphere, which enhances the overall aesthetic of the image.
Question: What type of company is the logo best suited for?

Figure 9: Another multi-turn question answering case generated by IPN (LLaMA-7b).



- 👤 Question: What is the man wearing in the image?
🤖 The man is wearing a suit and tie.
👤 Question: What is the color scheme of the image?
🤖 The color scheme of the image is black and white.
👤 Question: What is the man's expression in the image?
🤖 The man's expression in the image is serious.
👤 Question: What could the image be used for?
🤖 The image can be used for a variety of purposes, such as advertising, marketing, or promotional materials. It can also be used for educational or informational purposes, such as in a textbook or presentation.
👤 Question: Is the man wearing gloves in the image?
🤖 Yes, the man is wearing gloves in the image.

Figure 10: A multi-turn question answering case generated by IPN (LLaMA-13b).

C In-detail Image Description



BLIP-2: a lake surrounded by mountains and flowers

LLaVA (Vicuna-7b): The image features a serene lake surrounded by mountains and hills, with a large island in the middle of the lake. The lake is calm and blue, reflecting the beautiful landscape. **There are several small boats scattered across the lake, some near the shore and others further away.** In the background, there are mountains of various sizes, adding to the picturesque scenery. The scene is a perfect representation of a peaceful, natural environment that one can find in the mountains.

IPN (LLaMA-7b)The image captures a beautiful scene of a lush green forest with tall trees and a river flowing through the area. The river runs through the forest, creating a peaceful and tranquil atmosphere. There are several trees in the scene, some of which are positioned closer to the river, while others are further away from the water. The tall trees create a dense canopy, providing shade and enhancing the natural beauty of the area. In addition to the trees and the river, there are a few rocks scattered throughout the forest.These rocks add to the natural beauty of the scene, creating a picturesque landscape. The overall atmosphere of the scene evokes a sense of serenity and tranquility, making it a perfect place to relax and unwind.

IPN (LLaMA-13b): The image shows a breathtaking view of a mountainous landscape with a river flowing through it. The landscape is filled with greenery, creating a lush and vibrant scene. The river appears to be winding its way through the mountains, adding to the scenic beauty of the area. The image captures the tranquility and serenity of the natural environment, inviting the viewer to take a moment to admire the stunning view.

IPN (Bloomz-7b1): The image depicts a beautiful, serene scene with a large body of water, surrounded by lush green foliage. The water is surrounded by mountains, creating a dramatic backdrop for the scene. There is a small island in the middle of the body of water, giving it a sense of seclusion and isolation. **In the middle of the body of water, there is a small boat with two people on board. They seem to be enjoying a peaceful day on the water, gazing at the beautiful scenery. The boat is positioned in the middle of the body of water, creating a sense of balance and harmony in the scene.** (Hallucination)

Figure 11: An in-detail description case generated by several models.



BLIP-2: an empty room with a fan in it

LLaVA (Vicuna-7b): The image features a large, dark room with a doorway that has a white frame. The doorway is open, leading to a hallway that appears to be empty. Inside the room, there is a brown, rusted metal cabinet that stands out in the space. **In addition to the cabinet, there are several books scattered throughout the room, adding to the overall cluttered appearance. The books can be seen on the floor, as well as on various surfaces within the room. The room seems to be in a state of disarray, with the open doorway leading to the equally mysterious hallway beyond.** (Hallucination)

IPN (LLaMA-13b): The image depicts a room with a white wall and a door on the left side. The room appears to be empty, with no furniture or people present. The door appears to be closed, suggesting that the room is not currently in use.

IPN (LLaMA-7b):The image depicts a room with a large wooden desk placed in the center of the space. The desk is covered with various items, including a clock, a book, a pencil, and a calculator. **There is also a vase placed on the desk, adding a touch of elegance to the room. In addition to the desk, there are two chairs placed in the room, one on each side of the desk. The room appears to be a classroom or an office, as there are several books placed on the desk and around the room.**(Hallucination)

IPN (Bloomz-7b1): The image depicts a small room filled with various pieces of furniture, including a table, a chair, and a bed. The furniture appears to be in good condition, with the table, chair, and bed all in good condition. **The room is filled with a variety of items, including a bookshelf, a lamp, and a mirror. The bookshelf is placed in the corner of the room, while the lamp and mirror are positioned in the middle of the room. There is also a small painting on the wall of the room, which adds a touch of color to the scene. Overall, the room appears to be in good condition and ready for use.**(Hallucination)

Figure 12: Another case of generated in-detail description.