Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models

Xin He, **Longhui Wei**[†], **Lingxi Xie** and **Qi Tian**
Huawei Inc.
{whut.hexin, weilh2568, 198808xc}@gmail.com, tian.qi1@huawei.com

Abstract

Multimodal Large Language Models (MLLMs) are experiencing rapid growth, yielding a plethora of novel works recently. The prevailing trend involves adopting data-driven methodologies, wherein diverse instruction-following datasets were collected. However, these approaches always face the challenge of limited visual perception capabilities, as they solely utilizing CLIP-like encoders to extract visual information from inputs. Though these encoders are pre-trained on billions of image-text pairs, they still grapple with the information loss dilemma, given that textual captions only partially capture the contents depicted in images. To address this limitation, this paper proposes to improve the visual perception ability of MLLMs through a mixture-of-experts knowledge enhancement mechanism. Specifically, this work introduces a novel method that incorporates multi-task encoders and existing visual tools into the MLLMs training and inference pipeline, aiming to provide a more comprehensive summarization of visual inputs. Extensive experiments have evaluated its effectiveness of advancing MLLMs, showcasing improved visual perception capability achieved through the integration of visual experts.

1 Introduction

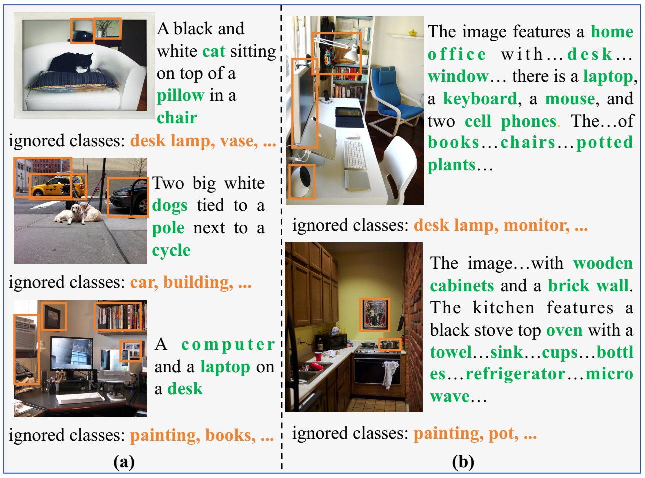
A black and white cat sitting on top of a pillow in a chair
ignored classes: desk lamp, vase, ...
Two big white dogs tied to a pole next to a cycle
ignored classes: car, building, ...
A computer and a laptop on a desk
ignored classes: painting, books, ...
(a)
The image features a home office with…desk… window… there is a laptop, a keyboard, a mouse, and two cell phones. The…of books…chairs…potted plants…
ignored classes: desk lamp, monitor, ...
The image…with wooden cabinets and a brick wall. The kitchen features a black stove top oven with a towel…sink…cups…bottles…refrigerator…microwave…
ignored classes: painting, pot, ...
(b)

Figure 1: Examples from public image-text pairs. (a) Examples from COCO Caption [Chen *et al.*, 2015]. (b) Examples from LLaVA-Instruct-150K [Liu *et al.*, 2023c]. The short textual captions in (a) only describe parts of the corresponding images. The captions in (b) are more informative but still cannot describe all the content inside these images. The orange boxes inside these images indicate objects that are missed in the captions.

Recently, the development of large language models (LLMs) [Chung *et al.*, 2022; Touvron *et al.*, 2023a; Touvron *et al.*, 2023b] has notably propelled advancements in artificial general intelligence. Various domains within artificial intelligence have actively embraced LLMs to enhance their performances across different tasks [Liu *et al.*, 2023c; Chen *et al.*, 2023; Hong *et al.*, 2023]. The multimodal dialogue field is no exception, witnessing a surge in the development of multimodal large language models (MLLMs) in recent months [Liu *et al.*, 2023c; Zhu *et al.*, 2023; Dai *et al.*, 2023; Ye *et al.*, 2023b; Bai *et al.*, 2023]. These works commonly insert visual encoders into LLMs, followed by fine-tuning a light-weight projection network to project extracted visual information into the language latent space.

While recent advancements have notably enhanced the performances of downstream multimodal dialogue tasks [Goyal *et al.*, 2017; Singh *et al.*, 2019; Mishra *et al.*, 2019; Masry *et al.*, 2022; Shah *et al.*, 2019], these improvements primarily stem from the collection of instruction data in various formats [Liu *et al.*, 2023c; Liu *et al.*, 2023a; Zhu *et al.*, 2023; Ye *et al.*, 2023a; Chen *et al.*, 2023]. Pioneering works such as MiniGPT-4 [Zhu *et al.*, 2023] and LLaVA [Liu *et al.*, 2023c] introduced an automatic mechanism for generating general multimodal instruction data, leveraging the capabilities of ChatGPT [OpenAI, 2023]. By subsequently fine-tuning MLLMs with the generated data, these approaches have achieved substantial enhancements in general multimodal tasks. Additionally, mPLUG-DocOwl [Ye *et al.*, 2023a] targets to amass instruction data related to documents, specifically enhancing the performance of MLLMs in document understanding tasks [Masry *et al.*, 2022; Mishra *et al.*, 2019;

[†] Corresponding author

Mathew *et al.*, 2021; Berant *et al.*, 2019]. Shikra [Chen *et al.*, 2023], on the other hand, proposed to collect referring expression pairs and fine-tune MLLMs on these pairs, thereby strengthening the models' ability to handle the referential dialogue tasks [Kazemzadeh *et al.*, 2014]. Furthermore, Instruct-BLIP [Dai *et al.*, 2023] and other related works [Bai *et al.*, 2023; Wang *et al.*, 2023] have proposed to assemble various multimodal datasets with distinct instruction templates. Subsequent fine-tuning of MLLMs on these consolidated datasets has proven instrumental in significantly improving their performances across different multimodal tasks.

As outlined above, while prior works have demonstrated advantages across various multimodal dialogue scenarios, they predominantly focus on collecting different types of instruction data, sharing a similar learning framework. Specifically, these works consistently employ a light-weight projection module (*e.g.*, Q-Former in BLIP-2 [Li *et al.*, 2023]) to project visual information, extracted by CLIP-like encoders (e.g., EVA-CLIP [Sun *et al.*, 2023]), into the language latent space. Given that the CLIP-like encoders cannot comprehensively describe the entirety of visual inputs (for them pre-trained with short textual captions, as shown in Fig.1(a)), MLLMs grapple with the visual information loss dilemma, which further affects the final performances. Moreover, though the detailed instruction data generated in LLaVA [Liu *et al.*, 2023c] or other works [Zhu *et al.*, 2023; Chen *et al.*, 2023] can alleviate the above problem to some extent, there are still lots of details that cannot be fully described (as shown in Fig. 1(b)). To address this challenge, there is a need for novel strategies that transcend the existing learning frameworks, enabling a more nuanced and accurate representation of visual inputs in MLLMs.

Inspired by the above, this paper explores MLLMs from the perspective of visual perception ability enhancement. Consequently, we introduce a simple but effective visual information learning framework, referred to as Incorporating Visual Experts (IVE), designed to augment the visual perception capabilities of MLLMs through aggregating available visual information extracted by specific experts. Specifically, IVE mainly involves two additional modules, *i.e.*, multi-task encoders and structural knowledge enhancement, for comprehensively describing the visual inputs. The multi-task encoders module integrates three auxiliary encoders, namely the low-level information encoder and the document-related information encoder, alongside with a CLIP-like encoder for semantics extraction. The above fusion aims to provide a more comprehensive description of visual inputs within the latent embedding space. The synergistic combination of these encoders facilitates a more nuanced understanding of the visual inputs. The structural knowledge enhancement module mainly utilizes specific visual tools to extract structural data (e.g., the categories and locations of instances or textual information inside images). These structural data will serve as prior knowledge and then be cooperated with the extracted latent embeddings fed into LLMs.

The introduced IVE is easy to implement, and its effectiveness has been substantiated through comprehensive experiments across various multimodal tasks. In general multimodal dialogue scenarios [Goyal *et al.*, 2017; Marino *et al.*, 2019], IVE excels in recognizing the intrinsic content of input images, thereby generating more accurate responses to each query in comparison to recent works [Liu *et al.*, 2023c; Zhu *et al.*, 2023]. Furthermore, when applied to specific multimodal dialogue tasks such as DocVQA [Mathew *et al.*, 2021], IVE shows competitive results compared with mPLUG-DocOwl [Ye *et al.*, 2023a], the recent method designed specifically for document analysis tasks. The above experiments further demonstrate the excellent visual perception ability achieved through the proposed integration strategy of visual experts.

2 Related Work

2.1 Vision-and-Language Pre-training

Recent multimodal large language models (MLLMs) [Liu *et al.*, 2023c; Zhu *et al.*, 2023; Dai *et al.*, 2023; Ye *et al.*, 2023a; Bai *et al.*, 2023; Ye *et al.*, 2023b] are commonly built on vision-and-language pre-training models (VLPs) [Chen *et al.*, 2020; Radford *et al.*, 2021; Li *et al.*, 2022], therefore this paper first revisits the development of VLPs before introducing MLLMs. The predominant VLP approaches can be broadly categorized into two frameworks: the one-stream framework [Chen *et al.*, 2020] and the two-stream framework [Radford *et al.*, 2021; Li *et al.*, 2022]. Methods within the one-stream framework [Chen *et al.*, 2020] typically employ a single transformer architecture to process both text and image data, incorporating various designs of loss functions. In contrast, the two-stream framework involves the independent extraction of modality information using distinct backbones. For efficiency, current MLLMs [Zhu *et al.*, 2023; Dai *et al.*, 2023; Ye *et al.*, 2023b] predominantly leverage the visual module of two-stream methods to encode the latent embeddings of visual inputs.

2.2 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have garnered considerable attention from both academia and industries, with a surge in novel works emerging recently [Liu *et al.*, 2023c; Dai *et al.*, 2023; Ye *et al.*, 2023b; Bai *et al.*, 2023; Huang *et al.*, 2024; Zong *et al.*, 2024; Xuan *et al.*, 2024]. A common framework underpins most of these works, featuring CLIP-like encoders responsible for extracting information from visual inputs, an abstractor summarizing the extracted information with few tokens, a light-weight layer further projecting the summarized information into the language latent space and a pre-trained large language model handling user questions in the context of the above extracted visual information. Despite their similar architectures, these works demonstrate versatility in addressing various multimodal dialogue tasks through training on distinct types of instruction data. For instance, LLaVA [Liu *et al.*, 2023c] excels in generating detailed answers for generic images with training on comprehensive instruction data. On the other hand, mPLUG-DocOwl [Ye *et al.*, 2023a] achieves significant improvements in the performance of MLLMs on document analysis tasks by training on document-related instruction data. Shikra [Chen *et al.*, 2023] enhances the model's capability in handling referring questions by training on referring expression pairs.

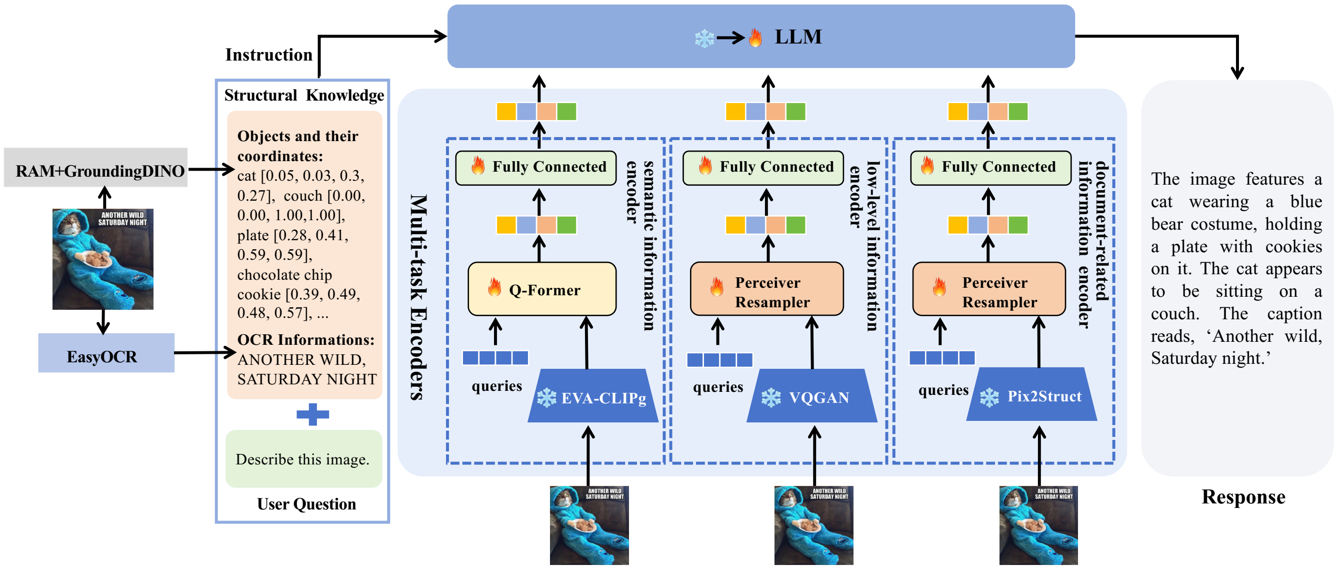
LLM
Instruction
Structural Knowledge
Objects and their coordinates:
cat [0.05, 0.03, 0.3, 0.27], couch [0.00, 0.00, 1.00,1.00], plate [0.28, 0.41, 0.59, 0.59], chocolate chip cookie [0.39, 0.49, 0.48, 0.57], ...
OCR Informations:
ANOTHER WILD, SATURDAY NIGHT
Describe this image.
User Question
RAM+GroundingDINO
EasyOCR
ANOTHER WILD SATURDAY NIGHT
Multi-task Encoders
Fully Connected
semantic information encoder
Q-Former
queries
EVA-CLIPg
low-level information encoder
Perceiver Resampler
VQGAN
document-related information encoder
Pix2Struct
The image features a cat wearing a blue bear costume, holding a plate with cookies on it. The cat appears to be sitting on a couch. The caption reads, 'Another wild, Saturday night.'
Response

Figure 2: The illustrations of IVE. Two modules, *i.e.*, the multi-task encoders and structural knowledge enhancement, are specifically designed in the framework. The multi-task encoders integrates multiple types of complementary encoders to collaboratively capture the latent information within visual inputs, and the structural knowledge enhancement module utilizes specific visual tools to detect the instances and textual information inside images as the prior knowledge fed into the large language model.

Although these works yield remarkable results, they remain constrained by the limited perception ability of CLIP-like encoders. In contrast to previous approaches, this work takes a novel perspective by focusing on enhancing the visual perception ability of MLLMs. The proposed approach involves aggregating available visual experts to provide a more comprehensive description of visual inputs, aiming to overcome the constraints imposed by the existing limitations in visual perception ability.

3 Our Approach

3.1 Preliminaries

Generally, the multimodal large language models (MLLMs) [Liu *et al.*, 2023c; Dai *et al.*, 2023; Ye *et al.*, 2023b; Bai *et al.*, 2023] are usually composed of three modules, *i.e.*, the visual perception module, the light-weight projection module, and the large language model, respectively. Specifically, the visual perception module extracts the inside contents from visual inputs and then the light-weight projection module projects the above visual information into the language latent space. The large language model receives the projected visual information and generates textual responses for each user query. Therefore, given the visual input as x_i, the query as q_i, the visual perception module as $\mathrm{F}_{\mathrm{vis}}(\cdot)$, the light-weight projection module as $\mathrm{F}_{\mathrm{proj}}(\cdot)$ and the large language model as $\mathrm{LLM}(\cdot)$, the process of generating response in MLLMs can be formulated as:

$$\mathrm{Response}_{q_i:x_i} = \mathrm{LLM}(\mathrm{F}_{\mathrm{proj}}(\mathrm{F}_{\mathrm{vis}}(x_i)), q_i), \tag{1}$$

where $\mathrm{Response}_{q_i:x_i}$ denotes the generated response for the query q_i based on the visual input x_i.

Limited by the computing and data resources, most current MLLMs directly utilize well-trained large language models, such as Flan-T5 [Chung *et al.*, 2022] and LLaMA [Touvron *et al.*, 2023a], as the encyclopedia to answer the given question. Therefore, the key for MLLMs lies in how to properly summarize the information of visual inputs into language space. Currently, most MLLMs [Bai *et al.*, 2023; Zhu *et al.*, 2023; Liu *et al.*, 2023c; Dai *et al.*, 2023; Ye *et al.*, 2023b; Ye *et al.*, 2023a] usually utilize CLIP-like encoders to extract the visual information, and then fine-tune a light-weight projection network with the collected instruction-following data to project extracted visual information into language latent space. Though extensive experiments have validated its effectiveness, the descriptions of visual inputs extracted by CLIP-like encoders are still not enough. As said ”a picture is worth a thousand words”, the CLIP-like encoders can only extract coarse semantic features inside each image in spite of their training on the billions of image-text pairs. To facilitate the above information loss dilemma, this paper proposes to incorporate visual experts in MLLMs, for comprehensively summarizing the visual contents of inputs. Details of the proposed approach will be described carefully in the next.

3.2 Incorporating Visual Experts into MLLMs

Different from previous works [Zhu *et al.*, 2023; Liu *et al.*, 2023c; Ye *et al.*, 2023b; Bai *et al.*, 2023; Chen *et al.*, 2023], this paper improves the visual perception ability of MLLMs from the perceptive of knowledge enhancement, and thus proposes a simple but effective framework with primarily **I**ncorporating different types of **V**isual **E**xperts into the current MLLMs, referred as IVE. As shown in Fig. 2, the visual perception within IVE relies on two pivotal modules: the multi-task encoders and structural knowledge enhancement module. The multi-task encoders is designed to amalgamating various types of latent visual information extracted by multiple visual encoders. This integration improves its comprehensiveness in the view of latent embedding of visual inputs. Additionally, the structural knowledge enhancement module is crafted to leverage visual tools, such as OCR tools [JaidedAI, 2020] and object detectors [Zhang *et al.*, 2023; Liu *et al.*, 2023d], to extract prior knowledge from images. This extracted knowledge is then treated as visual prompts and incorporated into the large language model alongside the previously fused latent embeddings. Through the above cooperative modules, IVE can comprehensively encode the contents of visual inputs from diverse perspectives, thereby enhancing the visual perception ability of MLLMs.

Multi-task Encoders. The majority of current MLLMs commonly rely on CLIP-like encoders for extracting semantic information from visual inputs. However, the limited perception ability associated with this strategy restricts their performances across various multimodal dialogue scenes. In contrast, IVE seeks to resolve this limitation by integrating multiple types of complementary encoders to collaboratively capture the latent information within visual inputs. As depicted in Fig. 2, three types of encoders are employed: the semantic information encoder, the low-level information encoder, and the document-related information encoder, each contributing distinct perspectives to the overall understanding of visual content.

The semantic information encoder is designed to extract the semantics from visual inputs and subsequently project them into the language embedding space. Consistent with prevalent methodologies [Chen *et al.*, 2023; Liu *et al.*, 2023c; Li *et al.*, 2023; Zhu *et al.*, 2023; Ye *et al.*, 2023b], IVE adopts the CLIP-like encoder proposed in BLIP-2 [Li *et al.*, 2023], where EVA-CLIPg [Sun *et al.*, 2023] is initially employed to extract visual information, followed by the Q-former [Li *et al.*, 2023] designed to condense this information into a concise representation using a few tokens. Leveraging extensive training with abundant image-text pairs, this encoder generates embeddings adept at capturing the global semantic information of each visual input. The process of semantic feature extraction can thus be presented as follows:

$$\mathrm{F_s}(x_i) = \mathrm{CrossAtt_Q}(\mathrm{Enc_{eva}}(x_i), \{T_0, T_1, ..., T_m\}), \quad (2)$$

where $\mathrm{Enc_{eva}}$ denotes the visual encoder of EVA-CLIPg [Sun *et al.*, 2023], $\mathrm{CrossAtt_Q}$ represents the operations in Q-Former [Li *et al.*, 2023], $\{T_0, T_1, ..., T_m\}$ denotes the query tokens and m is the sum of query tokens, respectively.

Given the brevity of captions that only provide a coarse description of the global semantics within each image, the semantic information extracted by Eq. (2) is apparently insufficient. To enhance the richness of detailed information within the extracted latent embedding, a low-level information extractor is introduced as the supplement. This paper adopts the encoder from VQGAN [Esser *et al.*, 2021] as the corresponding low-level information extractor, which can encode images into latent embedding and then reconstruct them into original images with the corresponding decoder. However, directly integrating the extracted embedding into MLLMs is costly because of its high dimensionality. Following Flamingo [Alayrac *et al.*, 2022], we also utilize several query tokens (32 tokens) to summarize this latent embedding with a 3-layer Perceiver Resampler [Alayrac *et al.*, 2022], and the resultant tokens are then considered as low-level latent embedding. Consequently, the process of low-level information extraction can be formulated as:

$$\mathrm{F_l}(x_i) = \mathrm{CrossAtt_{PR}}(\mathrm{Enc_{vqgan}}(x_i), \{T_0, T_1, ..., T_n\}), \quad (3)$$

where $\mathrm{Enc_{vqgan}}$ denotes the pre-trained encoder of VQGAN [Esser *et al.*, 2021], $\mathrm{CrossAtt_{PR}}$ represents the operations in Perceiver Resampler [Alayrac *et al.*, 2022] and n represents the sum of query tokens for low-level information, respectively.

While the aforementioned low-level information extractor contributes additional details upon the semantic embedding, it's noteworthy that both are trained on general images and may lack specificity for certain types, such as the document image. To address this, a document-related information encoder is incorporated into the latent embedding learning framework. In our framework, Pix2Struct-Large [Lee *et al.*, 2023], a recent state-of-the-art approach in document analysis tasks, is employed for this purpose. Similar to the low-level information encoder, 64 query tokens are employed to succinctly summarize the extracted document-related information using a 6-layer Perceiver Resampler [Alayrac *et al.*, 2022]. Generally, the process of document-related information extraction can be formulated as:

$$\mathrm{F_d}(x_i) = \mathrm{CrossAtt_{PR}}(\mathrm{Enc_{pix}}(x_i), \{T_0, T_1, ..., T_k\}), \quad (4)$$

where $\mathrm{Enc_{pix}}$ denotes the pre-trained encoder of Pix2Struct-Large [Lee *et al.*, 2023] and k represents the sum of query tokens for document-related information.

Consequently, the final fused latent embeddings of each image in IVE can be formulated:

$$f^l_{x_i} = [\mathrm{F^s_{proj}}(\mathrm{F_s}(x_i)); \mathrm{F^l_{proj}}(\mathrm{F_l}(x_i)); \mathrm{F^d_{proj}}(\mathrm{F_d}(x_i))], \quad (5)$$

where $\mathrm{F^s_{proj}}$, $\mathrm{F^l_{proj}}$ and $\mathrm{F^d_{proj}}$ represents the linear projection layer for projecting the semantic information, low-level informaion and documented-related information into a common language feature space, respectively.

Structural Knowledge Enhancement. In view of that query tokens for each extractor undergo end-to-end training, ensuring that the summarized embeddings encompass the entirety of visual input remains a challenge. Thereby, this paper further introduces a structural knowledge enhancement module to explicitly extract structural data within each image using specific visual tools. Finally, these data are subsequently treated as prior knowledge and fed into the large language model alongside the fused latent embeddings.

Typically, human observation of an image involves first identifying the objects (their categories and locations) or textual information within this image. Drawing inspiration from this human cognitive process, the structural knowledge enhancement module is designed to extract three types of information: the category and localization of instances, together with textual content, respectively. We first utilize two specific visual tools (*i.e.*, RAM [Zhang *et al.*, 2023] and Grounding DINO [Liu *et al.*, 2023d]) to recognize and localize the objects inside each image. Furthermore, we utilize EasyOCR [JaidedAI, 2020] to detect the contained textual information of each visual input. Therefore, thanks to the above visual tools, most instances $[(c_0, x_0^0, y_0^0, x_0^1, y_0^1), ..., (c_q, x_q^0, y_q^0, x_q^1, y_q^1)]$ and textual information $[t_0, t_1, ..., t_o]$ inside each image can be detected, where c_i denotes the category of the detected i-th instance, $(x_i^0, y_i^0, x_i^1, y_i^1)$ represents the corresponding bounding boxes, t_j means the detected j-th visual text segment, q and o are the sum of detected instances or textual segments, respectively. Thereby, the final extracted structural knowledge can be formulated as:

$$f_{x_i}^s = [(c_0, x_0^0, y_0^0, x_0^1, y_0^1), ...,(c_q, x_q^0, y_q^0, x_q^1, y_q^1); t_0, t_1, ..., t_o], \tag{6}$$

To better align with LLM, we design the template in which inserting the extracted structural knowledge. The details of this template are shown in Appendix A of the supplemental materials.

Difference with previous works. While extant works [Gao *et al.*, 2023; Shen *et al.*, 2023], exemplified by LLaMA-Adapter v2 [Gao *et al.*, 2023], have explored the integration of visual tools to extract structural knowledge with the aim of augmenting the visual perceptual capabilities of MLLMs, it is notable that these approaches have predominantly restricted the deployment of visual tools solely into the inference stage. In contrast, the proposed IVE is meticulously crafted to harness structural knowledge throughout both the training and inference phases of MLLMs. This strategic design of IVE serves the dual purpose of mitigating the inherent noise introduced by the visual tools and comprehensively capitalizing on the helpful cues they provide.

3.3 Training Pipeline

Once the latent embeddings and structural knowledge are available, we feed them into a well-trained large-scale language model (LLaMA2-chat (7B) [Touvron *et al.*, 2023b], utilized in this work) and conduct the overall training, which makes LLM better handle these prompts while ignoring the inevitable noises. Following previous works [Bai *et al.*, 2023; Dai *et al.*, 2023], we reorganize several public multimodal datasets [Marino *et al.*, 2019; Masry *et al.*, 2022; Mathew *et al.*, 2021; Liu *et al.*, 2023c], and conduct supervised fine-tuning on them. Overall, the training process goes through three stages: pretraining, multi-task tuning, and specific fine-tuning. In the pretraining stage, we primarily utilize weakly labeled image-text pairs to train the alignment module in the semantic information encoder. The multi-task tuning stage involves training on various multimodal instruction datasets. Subsequently, in the specific fine-tuning stage, we fine-tune the model on the training set of specific datasets to better adapt to their unique characteristics. Details of each training process are provided in Appendix B of the supplemental materials.

4 Experiments

4.1 Datasets

Training Dataset. The entire training pipeline comprises three stages. In the pre-training stage, about 300M image-text pairs crawled from the Internet [Li *et al.*, 2022] are initially utilized to train Q-Former. Subsequently, the LLaVA-CC3M-Pretrain-595K from LLaVA [Liu *et al.*, 2023c] is employed to further train Q-Former and the projection layer. In the multi-task tuning stage, following previous work [Bai *et al.*, 2023], multi-task datasets are combined to jointly guide the further training of IVE, including several general VQA datasets (*e.g.*, VQAv2 [Goyal *et al.*, 2017]), OCR-related VQA datasets (*e.g.*, OCRVQA [Mishra *et al.*, 2019]), document-related VQA datasets (*e.g.*, DocVQA [Mathew *et al.*, 2021]), grounding datasets (*e.g.*, RefCOCO [Kazemzadeh *et al.*, 2014]), image captioning datasets (*e.g.*, COCO Caption [Chen *et al.*, 2015]), and multimodal instruction datasets (*e.g.*, LLaVA-Instruct-150K [Liu *et al.*, 2023c]). The statistics of the used training data in the mutl-task tuning stage are presented in Appendix C of the supplemental materials. In the fine-tuning stage, further fine-tuning is conducted on the training set of specific datasets individually to fit their unique characteristics, thus achieving better performances on specific tasks.

Evaluation Dataset. The evaluations cover general scene recognition, character recognition, chart and document analysis, as well as other multimodal dialogue tasks. To this end, the VQAv2 test set [Goyal *et al.*, 2017], OKVQA test set [Marino *et al.*, 2019], TextVQA validation set [Singh *et al.*, 2019], OCRVQA test set [Mishra *et al.*, 2019], DocVQA validation set [Mathew *et al.*, 2021], ChartQA test set [Masry *et al.*, 2022], WTQ test set [Berant *et al.*, 2019], and MME Benchmark [Fu *et al.*, 2023] are chosen for the evaluations.

4.2 Implementation Details

The overall training process of IVE includes three stages. In the pre-training stage, only the Q-Former and the projection layer of the semantic information encoder are trainable, while the other two encoders are temporarily removed. Moreover, the parameters of other modules remain frozen. The input resolution for the semantic information encoder is set as 224×224. When training with the 300M image-text pairs [Li *et al.*, 2022], the training encompasses only 1 epoch, and a global batch size of 2048. While training with the LLaVA-CC3M-Pretrain-595K [Liu *et al.*, 2023c], the training encompasses 5 epochs. The learning rate in this stage employs a cosine warm-up strategy (2000 steps), with a maximum learning rate of $1e$-4, and a minimum learning rate of $1e$-6.

In the multi-task tuning and specific fine-tuning stage, the language model undergoes tuning using LoRA [Hu *et al.*, 2021] with the hyper-parameters of rank=64. The Q-Former, Perceiver Resampler, and their corresponding projection layers are trainable, while the parameters of other modules remain frozen. The input resolution for the semantic information encoder is increased to 448×448, while the low-level information encoder is configured with the input resolution of 256×256. The input resolution of the document-related information encoder is set as 1024×1024. As for the learning rate, we employ a cosine warm-up strategy (500 steps), with a minimum learning rate of $1e$-6 and a maximum learning rate of $3e$-5 for the multi-task tuning stage, $1e$-5 for the specific fine-tuning stage. AdamW serves as the optimizer for all three training stages, with $\beta 1 = 0.9$, $\beta 2 = 0.98$, and the weight decay of 0.05, respectively.

Model	LLM	VQAv2	OKVQA	TextVQA	ChartQA	OCRVQA	WTQ	DocVQA
BLIP-2 [Li *et al.*, 2023]	13B	65.0	45.9	42.4	-	-	-	-
InstructBLIP [Dai *et al.*, 2023]	13B	-	-	50.7	-	-	-	-
Shikra [Chen *et al.*, 2023]	13B	77.4	47.2	-	-	-	-	-
mPLUG-DocOwl [Ye *et al.*, 2023a]	7B	-	-	52.6	57.4	-	26.9	62.2
Qwen-VL-Chat [Bai *et al.*, 2023]	7B	78.2	56.6	61.5	**66.3**	70.5	-	62.6
LLaVA-1.5 [Liu *et al.*, 2023b]	7B	78.5	-	58.2	-	-	-	-
mPLUG-Owl2 [Ye *et al.*, 2023b]	7B	**79.4**	57.7	58.2	-	-	-	-
SPHINX-Intern2 [Liu *et al.*, 2024]	7B	75.5	55.5	-	-	-	-	-
IVE(ours)	7B	78.8	**60.3**	**62.0**	65.3	**71.1**	**29.8**	**64.1**

Table 1: The direct-transfer results on VQA datasets.

Model	LLM	VQAv2	OKVQA	OCRVQA	ChartQA
BLIP2 [Li *et al.*, 2023]	13B	82.2	59.3	72.7	-
GIT2 [Wang *et al.*, 2022]	-	81.7	-	70.3	-
InstructBLIP [Dai *et al.*, 2023]	13B	-	62.1	73.3	-
CogVLM [Wang *et al.*, 2023]	7B	**84.7**	64.7	74.5	-
Pix2Struct-Large [Lee *et al.*, 2023]	-	-	-	71.3	58.6
IVE(ours)	7B	84.0	**65.2**	**74.9**	**68.3**

Table 2: The fine-tuning results on VQA datasets.

Model	LLM	Perception	Cognition
Qwen-VL-Chat [Bai *et al.*, 2023]	7B	1487.6	360.7
LLaVA-1.5 [Liu *et al.*, 2023b]	7B	**1510.7**	-
mPLUG-Owl2 [Ye *et al.*, 2023b]	7B	1450.2	313.2
SPHINX-Intern2 [Liu *et al.*, 2024]	7B	1260.4	294.6
IVE(Ours)	7B	1455.6	**384.1**

Table 3: The evaluations on MME Benchmark.

4.3 Direct-transfer performances on VQA Datasets

The VQA task entails multimodal large language models (MLLMs) answering questions based on both the visual inputs and user query. In this section, we conduct direct-transfer evaluations on multiple VQA benchmarks using the IVE model trained after multi-task tuning stage. We compare the proposed method with several state-of-the-arts, including Qwen-VL-Chat [Bai *et al.*, 2023], mPLUG-DocOwl [Ye *et al.*, 2023a], mPLUG-Owl2 [Ye *et al.*, 2023b], and LLaVA-1.5 [Liu *et al.*, 2023b]. The evaluation encompasses seven benchmarks: VQAv2 [Goyal *et al.*, 2017] and OKVQA [Marino *et al.*, 2019] for the general VQA task, TextVQA [Singh *et al.*, 2019] and OCRVQA [Mishra *et al.*, 2019] for the OCR VQA task, and ChartQA [Masry *et al.*, 2022], DocVQA [Mathew *et al.*, 2021], and WTQ [Berant *et al.*, 2019] for the document or chart VQA task, respectively. We employ the following prompt template for all evaluations on these datasets: "<Img>{latent embedding}</Img>{structural knowledge}{question}. Answer the question using a single word or phrase." In addition, as the object detection results of the chart and document images are usually useless, we design an automatic filtering mechanism to filter out the detection results of these images.

As presented in Tab. 1, our proposed method shows competitive performances when compared to recent approaches. Specifically, IVE achieves an accuracy of 60.3% on OKVQA [Marino *et al.*, 2019], which significantly surpasses the performance of recent state-of-the-art method (mPLUG-Owl2 [Ye *et al.*, 2023b], achieved with 57.7%). In TextVQA [Singh *et al.*, 2019] and OCRVQA [Mishra *et al.*, 2019] datasets, IVE achieves accuracies of 62.0% and 71.1%, outperforming Qwen-VL-Chat [Bai *et al.*, 2023] with 0.5% and 0.6%, respectively. As for the DocVQA [Mathew *et al.*, 2021] and WTQ [Berant *et al.*, 2019], IVE still achieves consistent improvements compared with recent approaches. More visualized examples have been shown in Appendix D of the supplemental materials.

Additionally, to demonstrate the robustness of our method, more validation about other widely-used VQA benchmarks with unseen images are shown in Appendix E of the supplemental materials.

4.4 Fine-tuning Performances on VQA Datasets

To compare our approach with different specific VQA methods, we assess the performance of IVE further fine-tuning on the VQAv2 [Goyal *et al.*, 2017], OKVQA [Marino *et al.*, 2019], OCRVQA [Mishra *et al.*, 2019], and ChartQA [Masry *et al.*, 2022]. We still employ the prompt template: "<Img>{latent embedding}</Img>{structural knowledge}{question}. Answer the question using a single word or phrase." during evaluation. The further fine-tuning results of IVE on these VQA datasets are shown in Tab. 2.

The experimental results demonstrate that IVE, following additional fine-tuning on specific datasets, achieves favorable improvements. Specifically, there are 5.2% and 4.9% improvements compared with the direct-transfer results on VQAv2 [Goyal *et al.*, 2017] and OKVQA [Marino *et al.*, 2019], respectively. Notably, in the tasks related to character and chart, IVE significantly outperforms the Pix2Struct-Large [Lee *et al.*, 2023] in OCRVQA [Mishra *et al.*, 2019] and ChartVQA [Masry *et al.*, 2022], with 3.6% and 9.7% improvements, respectively. Additionally, when compared to the recent state-of-the-art (CogVLM [Wang *et al.*, 2023]), IVE still shows competitive results.

Given that the MME Benchmark [Fu *et al.*, 2023] focuses on answering "yes/no" formats, we conduct further fine-tuning of our multi-task tuning model using a mixed dataset composed of VQAv2 [Goyal *et al.*, 2017] and LRV-Instruction [Liu *et al.*, 2023a]. Subsequently, we evaluate the model on the MME Benchmark. As demonstrated in Tab. 3, our method achieves the scores of 1455.6 and 384.1 in the perception and cognition task of MME Benchmark [Fu *et al.*, 2023], respectively. Compared with recent state-of-the-arts (mPLUG-Owl2 [Ye *et al.*, 2023b] and LLaVA-1.5 [Liu *et al.*,

Methods	VQAv2	OKVQA	TextVQA	ChartQA	OCRVQA	WTQ	DocVQA
semantic information encoder only	75.6	57.3	56.3	58.7	67.5	27.1	60.2
+ low-level information encoder	76.7	57.8	58.5	59.1	68.5	27.3	60.8
+ document-related information encoder	77.2	58.0	59.9	61.8	70.2	27.9	62.3
+ structural knowledge enhancement on Infer	76.2	57.6	60.4	63.3	70.6	28.6	62.7
+ structural knowledge enhancement on Train&Infer	**78.8**	**60.3**	**62.0**	**65.3**	**71.1**	**29.8**	**64.1**

Table 4: The ablation studies of each proposed module on VQA datasets.

2023b]), our IVE demonstrates superior stability.

4.5 Ablation Study

To better evaluate the effectiveness of each proposed module, we further conduct ablation studies with utilizing different combinations of proposed modules. All the ablations are conducted with training on multi-task tuning datasets and performing direct-transfer evaluations on different VQA test sets.

Effectiveness of Multi-task Encoders. To evaluate the individual contributions of each encoder within our multi-task encoders, three distinct experiments have been conducted. The initial experiment exclusively employs the semantic information encoder. Subsequently, the low-level information encoder and document-related information encoder are utilized progressively. As shown in Tab. 4, combing the semantic information encoder and the low-level information encoder leads to improvements across various datasets compared to only using the semantic information encoder. Further fusion with the document-related information encoder results in a significant improvement on document-related analysis tasks, with its performance on ChartVQA [Masry *et al.*, 2022] rising from 59.1% to 61.8% and DocVQA [Mathew *et al.*, 2021] rising from 60.8% to 62.3%, respectively. More visualized analysis have been present in Appendix D of the supplemental materials.

Effectiveness of Structural Knowledge Enhancement. To validate the effect of structural knowledge enhancement and compare the different impacts of integrating structural knowledge only in the inference phase or in both the training and inference phases, we further conduct two additional experiments built upon the multi-task encoders.

As shown in Tab. 4, while evaluating VQAv2 [Goyal *et al.*, 2017] and OKVQA [Marino *et al.*, 2019], the performances will decrease with incorporating structural knowledge solely during the inference phase. Conversely, integrating this structural knowledge during both the training and inference phases yields improved results across a spectrum of datasets. The above phenomenon demonstrates that the structural knowledge introduces inherent noises, negatively impacting the capability of MLLMs while it is directly utilized. However, when introducing these extracted knowledge during the training phase, the LLM is guided to autonomously discern pertinent information, thereby mitigating the adverse effects of noise. More visualized analysis and effectiveness validation about structural knowledge enhancement module are shown in Appendix D and Appendix F of the supplemental materials, respectively.

Moreover, to demonstrate that integrating structural knowledge during both training and inference phases can mitigate the disturbation of noises in knowledge rather than simply aligning prompt formats, we conduct additional experiments with fine-tuning on the sampled VQAv2 [Goyal *et al.*, 2017] dataset. Specifically, we replace the automatically detected results with the ground truth as finally utilized structural knowledge. Then, we conduct the comparisons with integrating structural knowledge only in the inference phase or in both training and inference phases. As shown in Tab. 5, utilizing the ground truth as structural knowledge and integrating it during both training and inference phases only achieves slight gains (0.5%) compared to the mechanism of integrating ground truth during the inference phase (1.6% gains). This observation demonstrates that our proposed method goes beyond simple prompt format alignment. Instead, it focuses on autonomously discerning and extracting pertinent information, thereby mitigating the adverse effects of noise.

Model	VQAv2
Multi-task Encoders	81.3
+ structural knowledge enhancement on Infer	82.9
+ structural knowledge enhancement on Train&Infer	83.4

Table 5: The ablation studies while regarding ground truth as the utilized structural knowledge.

5 Conclusion

This paper firstly reevaluates the existing limitations within current multimodal large language models(MLLMs), and points out that they always grapple with the information loss dilemma. To enhance the corresponding visual perception ability of MLLMs, this paper presents Incorporating Visual Experts(IVE), the first work to aggregate available visual information through a mixture-of-experts mechanism in both training and inference stages. Extensive experiments on a wide range of multimodal dialogue datasets have evaluated the effectiveness of IVE. Though the significant improvements achieved by IVE, the types of visual experts utilized in current pipeline are still limited. In addition, compared to the methods which use only a semantic encoder branch, IVE incurs an additional 979M parameters, which is far less than the total parameters of MLLMs. In the future, we aim to explore more efficient multi-encoder fusion strategies and develop a more effective information compression projector.

Contribution Statement

Xin He and Longhui Wei are co-first authors.

References

[Alayrac *et al.*, 2022] 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.

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

[Berant *et al.*, 2019] Jonathan Berant, Daniel Deutch, Amir Globerson, Tova Milo, and Tomer Wolfson. Explaining queries over web tables to non-experts. In *2019 IEEE 35th international conference on data engineering (ICDE)*, pages 1570–1573. IEEE, 2019.

[Chen *et al.*, 2015] 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.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

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

[Chung *et al.*, 2022] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[Esser *et al.*, 2021] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[Gao *et al.*, 2023] 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.

[Goyal *et al.*, 2017] 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.

[Hong *et al.*, 2023] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *arXiv preprint arXiv:2307.12981*, 2023.

[Hu *et al.*, 2021] 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.

[Huang *et al.*, 2024] Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: multimodal large language models can enhance visual robustness via denoising in-context learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 19973–20003, 2024.

[JaidedAI, 2020] JaidedAI. Easyocr. https://github.com/JaidedAI/EasyOCR, 2020.

[Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[Lee *et al.*, 2023] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023.

[Li *et al.*, 2022] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022.

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

[Liu *et al.*, 2023a] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1, 2023.

[Liu *et al.*, 2023b] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[Liu *et al.*, 2023c] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[Liu *et al.*, 2023d] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[Liu *et al.*, 2024] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

[Marino *et al.*, 2019] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[Masry *et al.*, 2022] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[Mathew *et al.*, 2021] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[Mishra *et al.*, 2019] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[OpenAI, 2023] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2023.

[Radford *et al.*, 2021] 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.

[Shah *et al.*, 2019] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.

[Shen *et al.*, 2023] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.

[Singh *et al.*, 2019] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[Touvron *et al.*, 2023a] 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.

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

[Wang *et al.*, 2022] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[Wang *et al.*, 2023] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[Xuan *et al.*, 2024] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13838–13848, 2024.

[Ye *et al.*, 2023a] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

[Ye *et al.*, 2023b] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[Zhang *et al.*, 2023] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023.

[Zhu *et al.*, 2023] 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.

[Zong *et al.*, 2024] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.