Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment

Yunxin Li[1], **Xinyu Chen**[1], **Baotian Hu**[1]*, **Haoyuan Shi**[1], **Min Zhang**[1]
[1]Harbin Institute of Technology, Shenzhen, China
liyunxin987@163.com
{hubaotian, zhangmin2021}@hit.edu.cn

Abstract

Evaluating and Rethinking the current landscape of Large Multimodal Models (LMMs), we observe that widely-used visual-language projection approaches (e.g., Q-former or MLP) focus on the alignment of image-text descriptions yet ignore the visual knowledge-dimension alignment, i.e., connecting visuals to their relevant knowledge. Visual knowledge plays a significant role in analyzing, inferring, and interpreting information from visuals, helping improve the accuracy of answers to knowledge-based visual questions. In this paper, we mainly explore improving LMMs with visual-language knowledge alignment, especially aimed at challenging knowledge-based visual question answering (VQA). To this end, we present a *Cognitive Visual-Language Mapper* (CVLM), which contains a pretrained Visual Knowledge Aligner (VKA) and a Fine-grained Knowledge Adapter (FKA) used in the multimodal instruction tuning stage. Specifically, we design the VKA based on the interaction between a small language model and a visual encoder, training it on collected image-knowledge pairs to achieve visual knowledge acquisition and projection. FKA is employed to distill the fine-grained visual knowledge of an image and inject it into Large Language Models (LLMs). We conduct extensive experiments on knowledge-based VQA benchmarks and experimental results show that CVLM significantly improves the performance of LMMs on knowledge-based VQA (average gain by 5.0%). Ablation studies also verify the effectiveness of VKA and FKA, respectively.[1]

1 Introduction

Recent Large Multimodal Models (LMMs) such as GPT-4V (OpenAI, 2023), Gemini (Team et al., 2023), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Liu et al., 2023b), LLaVA (Liu et al., 2023a), and many others, have achieved impressive performance in a variety of visual understanding and reasoning tasks, especially on Visual Question Answering (VQA) (Li et al., 2023e,g). Current open-source LMMs are usually constructed by combining pertained visual encoders and Large Language Models (LLMs), as depicted in the bottom part of Figure 1, where a visual mapping network (e.g., Q-former (Li et al., 2023c), Linear (Zhu et al., 2023), or MLP (Liu et al., 2023a; Li et al., 2023f)) is employed to project visual representations into the language space of LLMs. Although such LMMs have achieved powerful visual understanding capability similar to GPT-4V and Genimi on some image understanding tasks such as Image Captioning (Changpinyo et al., 2021), Visual Dialogue (Zhang et al., 2022; Chen et al., 2022), Visual Entailment (Xie et al., 2019; Do et al., 2020), and VQA (Antol et al., 2015), they often fall short of knowledge-based VQA, which necessitates relevant knowledge to answer these visual questions.

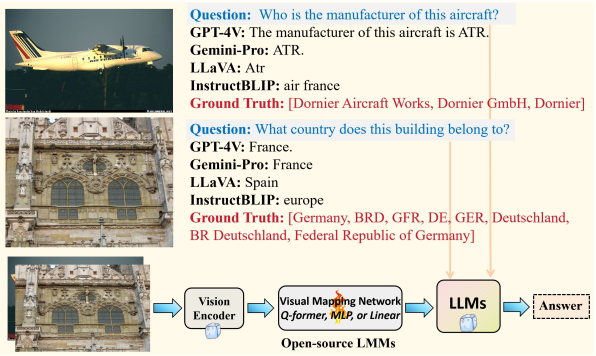

Figure 1: It illustrates the performance of LMMs on visual information-seeking questions. The bottom part shows the widely-used architecture of open-source LMMs, where the visual mapping network is usually pretrained on massive image-text captioning data. All LMMs including GPT-4V (Date: 2023.11.17) and Gemini-Pro make incorrect decisions.

* Corresponding author.
[1]Codes and Data are available at https://github.com/HITsz-TMG/Cognitive-Visual-Language-Mapper

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7615–7626

As the cases illustrated in Figure 1, these advanced LMMs (including GPT-4V and Gemini-Pro) can not give correct answers to simple visual information seeking questions: *Who is the manufacturer of this aircraft; What country does this building belong to?*.

In light of this, rethinking the construction process of LMMs (Wang et al., 2022; Li et al., 2023g; Zhu et al., 2023; Liu et al., 2023a; Koh et al., 2023) from the initial pretraining stages, we discover that these visual mapping networks trained on massive image-text captioning pairs simply transfer visual features to their language descriptions. They overlook the visual language knowledge-dimension alignment. i.e., connecting visuals to their relevant knowledge. As we know, visual knowledge (Collins and Olson, 2014) plays a pivotal role in the way humans understand and interact with the world. It extends beyond the mere ability to recognize and interpret visuals, incorporating an understanding of spatial relationships, patterns, and symbols, which are essential components of human cognition (Pinker, 1984; Cavanagh, 2011). Additionally, previous works also demonstrated that introducing visual knowledge (Lu et al., 2022b; Zhu et al., 2022; Li et al., 2023f,h) can improve the performance of pretrained language models on natural language understanding (Lu et al., 2022b) and open-ended text generation tasks (Zhu et al., 2022). Inspired by these insights, we focus on enhancing LMMs through the introduction of visual-language knowledge alignment, going beyond the conventional scope of visual-language integration.

To this end, we present a *Cognitive Visual-Language Mapper* (**CVLM**), which contains a pretrained Visual Knowledge Aligner (VKA) and a Fine-grained Knowledge Adapter (FKA). Specifically, we devise VKA based on a small language model that interacts with fine-grained image representation in each block. The output hidden states of VKA are fed into the LLM as the knowledge embedding tokens by a linear projection layer. To make VKA effectively capture image-relevant knowledge, we first train it on image-knowledge pairs (Srinivasan et al., 2021) collected from Wikipedia [2] via the next tokens prediction. Like Q-former (Li et al., 2023c) and prefix-tuning (Li and Liang, 2021), we only fine-tune some learnable query tokens and the linear layer to acquire fixed-length visual knowledge representation and convert it into the representation space of LLM. In addition, considering that visual objects contain fine-grained visual knowledge, we introduce FKA to gain comprehensive visual knowledge of an image and distill valuable visual knowledge from the whole knowledge representation sequence. The output knowledge vectors of FKA are injected into each layer of LLMs to realize in-depth interactions between LLMs and detailed visual knowledge. By doing so, CVLM is capable of connecting visuals to relevant knowledge, enabling LMMs to utilize them during multimodal understanding and generation.

To verify the effectiveness of CVLM, we conduct extensive experiments on image-centered, knowledge-based, and complex visual reasoning question-answering scenarios: VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), Infoseek (Chen et al., 2023a), TextVQA (Singh et al., 2019), and SeedBench (Li et al., 2023a). The experimental results show that CVLM significantly outperforms previous strong baselines such as LLaVA-v1.5. The ablation and case studies indicate that CVLM is capable of linking visual knowledge and improving performance on knowledge-intensive tasks via the introduced aligner and adapter.

Our contributions can be summarized as follows:

- We present a cognitive visual-language mapper to achieve visual-language knowledge alignment, which contains a pretrained visual knowledge aligner and a fine-grained knowledge adapter that is used to distil and inject valuable visual knowledge into LLMs.
- To the best of our knowledge, we are the first to explore the visual-language knowledge alignment during the pretraining and finetuning stages of LMMs, connecting visuals to their knowledge via CVLM.
- Experimental results indicate that CVLM significantly improves the performance of LMMs on knowledge-intensive VQA. The ablation studies also verify the effectiveness of VKA and FKA on specific knowledge-based VQA.

2 Related Work

Knowledge-based Visual Question Answering. Visual Question Answering (VQA) is a multidisciplinary field that combines vision and language

[2] https://en.wikipedia.org/wiki/Wikipedia:Images

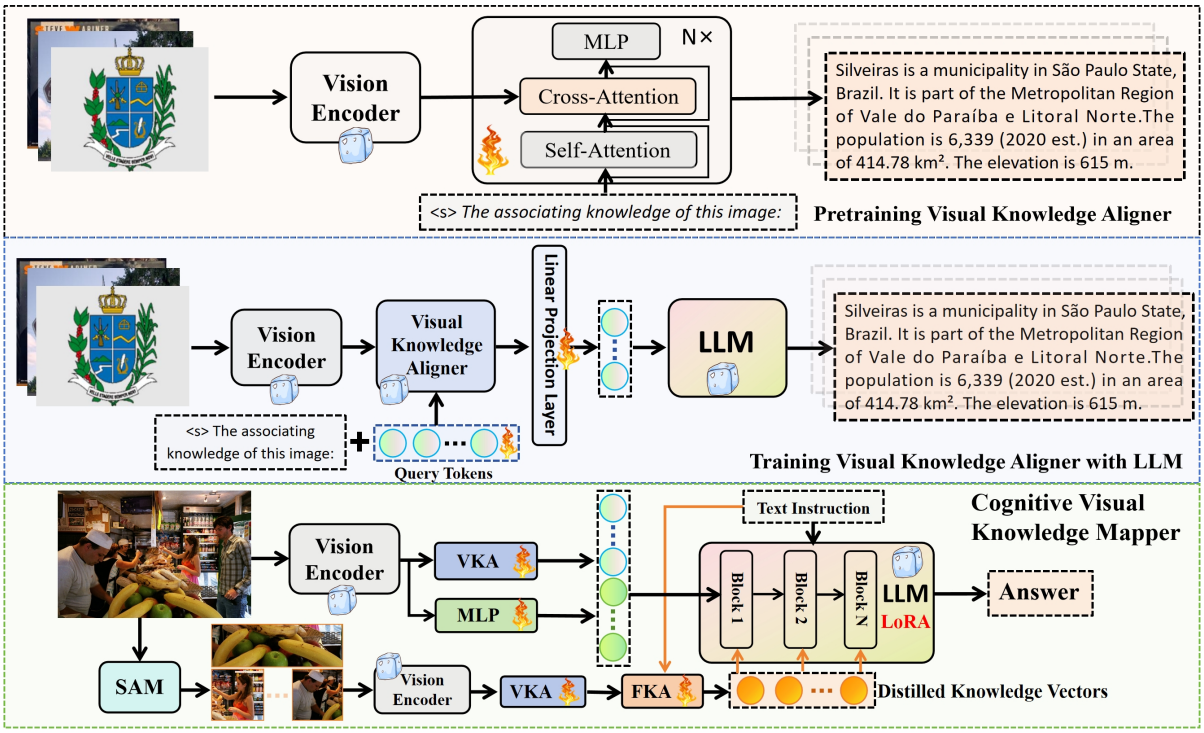

Figure 2: An overview of Cognitive Visual Knowledge Mapper. From top to bottom, it shows 1) Pretraining visual knowledge aligner, where we use a pretrained small language model to interact with image features via the cross attention module; 2) Training visual knowledge aligner with LLM, in which we realize visual knowledge alignment between vision encoder and LLM via the learnable query tokens and linear layer; 3) Overall architecture of CVLM, where we present the fine-grained visual knowledge adapter beyond common visual projection (MLP) and VKA.

processing to address queries about images. A recent development in this domain is knowledge-based VQA, which relies on external information for open-domain visual questions. The initial knowledge-based VQA datasets, KB-VQA (Wang et al., 2015) and FVQA (Wang et al., 2017) had limited knowledge requirements, referred to as "closed" knowledge. In contrast, S3VQA (Jain et al., 2021) and OK-VQA (Marino et al., 2019) datasets introduced questions demanding "open-domain" knowledge, incorporating widely recognized facts from diverse domains. INFOSEEK (Chen et al., 2023a), a recent Wikipedia-based VQA dataset, concentrated on fine-grained entity knowledge for open-domain information-seeking queries. As a result, datasets like OK-VQA and INFOSEEK, encompassing diverse knowledge categories, are ideal for assessing the performance of LMMs in open-domain VQA tasks. A-OKVQA (Schwenk et al., 2022) necessitates a broad foundation of common sense and worldly knowledge for answering questions, like QA on knowledge graph (Chen et al., 2023d,c).

Large Visual-Language Models. Recent advancements in foundational models for vision and language have led to the development of LMMs. In response to large model GPT-4 (OpenAI, 2023), several others have emerged, including GVT (Wang et al., 2023), MPlug (Ye et al., 2023), Macaw (Lyu et al., 2023), LMEye (Li et al., 2023g), and LLaVA (Liu et al., 2023a). These models have demonstrated strong performance across various visual-language tasks. Typically, these models utilize pretrained visual models to extract visual features, which are then integrated into the linguistic space of LMMs through a straightforward projection layer. This layer can be a Linear Layer (Merullo et al., 2023; Liu et al., 2023b; Li et al., 2023f) or a Q-former (Li et al., 2023c). Following this integration, similar to the supervised fine-tuning approach used for LLMs, these systems undergo refinement using diverse and high-quality multimodal instruction-following datasets (Liu et al., 2023b; Zhu et al., 2023; Ye et al., 2023). These datasets encompass both human-labeled data for downstream tasks like Visual Question Answering (VQA) and VCR (Antol et al., 2015; Zellers et al., 2019), as well as datasets automatically generated by GPT-4. Meanwhile, some multimodal benchmarks like MM-

Bench (Liu et al., 2023c) and SEED-Bench (Li et al., 2023a) have been established to evaluate advanced LMMs, and Li et al. (2023i) presents a comprehensive assessment of their performance in knowledge-intensive VQA scenarios.

3 Methodology

3.1 Overview

CVLM focuses on connecting visuals to relevant knowledge and injecting them into LLMs to perform multimodal instruction-following generation. The overview of CVLM is illustrated in Figure 2. Specifically, given an image I and text instruction $T = (t_1, t_2, ..., t_M)$, where t_M refers to the M th token of instruction, we initially utilize a visual encoder to generate representations of images. These representations are then mapped into the language space of LMMs using a Multilayer Perceptron (MLP). We integrate a Visual Knowledge Aligner (VKA) between the frozen visual encoder and LLMs to transform the visual knowledge into the language space of LLMs. Furthermore, acknowledging that image regions detailed visual knowledge, we introduce a Fine-Grained Visual Knowledge Adapter (FKA). This adapter is designed to extract valuable information from the intricate visual knowledge representations produced by the VKA. This distilled knowledge is subsequently incorporated into the LLMs. Through this methodology, we enable the LLMs to not only associate with but also utilize visual knowledge effectively, thereby facilitating multimodal generation in an end-to-end fashion.

3.2 Visual Knowledge Aligner

Firstly, we introduce a task-agnostic visual knowledge generator to realize associating relevant visual knowledge given an image. Specifically, we employ a pretrained visual encoder CLIP ViT-L/14 with inputting image size of 336*336 to gain the image representation sequence $\mathbf{h}_I = (h_g, h_1^I, ..., h_{576}^I)$, where h_g and h_{576}^I refers to the global feature of the image and 576 th patch representations. Then, as shown in the top part of Figure 2, we utilize a pretrained small autoregressive language model (OPT-1.3B) as the generator of visual knowledge, which interacts with the visual sequence h_I via adding the cross attention layer in each block. We train it on an amount of image-knowledge pairs via the next token prediction. These pairs are meticulously curated from Wikidata, ensuring a rich and diverse source of world knowledge information. This pretrained knowledge generator is capable of associating relevant knowledge based on an input image.

Afterward, as depicted in the middle part of Figure 2, we use the pretrained visual knowledge generator as the backbone to construct the whole visual-knowledge aligner like Q-former. Specifically, we add learnable tokens $\mathbf{h}_{\text{KQ}} = (h_{\text{KQ}}^1, ..., h_{\text{KQ}}^N)$, where N refers to the number of query tokens, which is joined with the knowledge prompt "*<s> The associating knowledge of this image*". We adopt a learnable linear projection layer to project the obtained features into the language space. The whole process could be presented in Eq. 1.

$$\begin{aligned} \mathbf{h}_{\text{AO}} &= \text{VKA}([\mathbf{h}_{\text{KP}}, \mathbf{h}_{\text{KQ}}], \mathbf{h}_I), \\ \mathbf{h}_{\text{KO}} &= \mathbf{W}^K \mathbf{h}_{\text{AO}} + \mathbf{b}, \end{aligned} \tag{1}$$

where $\mathbf{h}_{\text{AO}}$ and $\mathbf{h}_{\text{KP}}$ show the output of the pretrained visual knowledge generator and the word embeddings of knowledge prompt, respectively. $[,]$ refers to the sequence concatenation of two vectors. $\mathbf{W}^K \in \mathbf{R}^{d_K \times d_L}$ and $\mathbf{b} \in \mathbf{R}^{d_L}$ are the learnable parameters, where d_K and d_L represent the hidden state dimensions of visual knowledge aligner and LLM, respectively. $\mathbf{h}_{KO}$ will be fed into the language models with the original image representation $\mathbf{h}_{\text{IO}}$. It is gained by a learnable MLP trained on image-text captioning pairs, i.e., $\mathbf{h}_{\text{IO}} = \text{MLP}([h_1^I, ..., h_{576}^I])$.

The supervision signal remains the sequence of knowledge relevant to the image and the learning objective is the cross-entropy generation loss. By doing so, the designed knowledge aligner could connect visuals to their knowledge and project them into the LLMs.

3.3 Fine-grained Visual Knowledge Adapter

Considering that image regions (e.g., objects) also link to useful knowledge, we present a fine-grained visual knowledge adapter (FKA) to gain and distill detailed and useful visual knowledge, which is injected into each block in LLMs. Firstly, we obtain the fine-grained visual knowledge representations by using widely used segment anything tools named SAM (Kirillov et al., 2023) and VKA. Concretely, as shown in the bottom-left part of Figure 2, we use the SAM to obtain the image regions with their confidence scores and adopt the top five image objects, which could be denoted as $I_1, ..., I_5$. Then, we utilize vision encoder and VKA to obtain

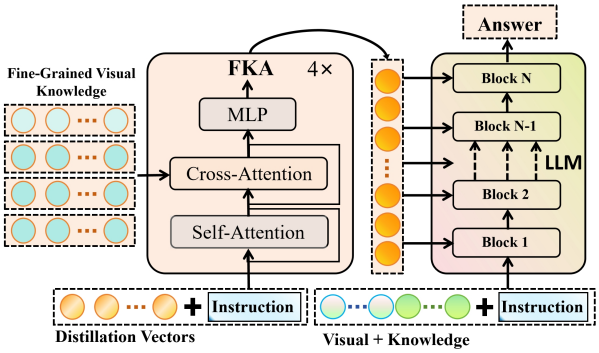

Figure 3: The detailed calculation process of fine-grained visual knowledge adapter, i.e., VKA shown in Figure 2. "Visual + Knowledge" indicates the representation concatenation of an image $\mathbf{h}_{\text{IO}}$ and its relevant knowledge projection $\mathbf{h}_{\text{KO}}$.

the fine-grained visual knowledge representations $\mathbf{h}_{I_1}, \mathbf{h}_{I_2}, ..., \mathbf{h}_{I_5}$, which are illustrated in Figure 3.

Subsequently, we employ a four-layer transformer decoder with learnable distillation vectors $\mathbf{h}_D = (\mathbf{h}_1^D, ..., \mathbf{h}_N^D)$, where N is the number of distill vectors. To gain useful knowledge, we splice text instruction T after $\mathbf{h}_D$ to distill instruction-relevant visual knowledge. The specific calculation progress of each FKA block is given in Eq. 2

$$\begin{aligned} \mathbf{h}_S^{l-1} &= \text{Self-A}(\mathbf{h}_D, \mathbf{h}_T) + \mathbf{h}_{\text{FKA}}^{l-1}, \\ \mathbf{h}_{cr} &= \text{Cross-A}(\mathbf{h}_S^{l-1}, [\mathbf{h}_{I_1}, \mathbf{h}_{I_2}, ..., \mathbf{h}_{I_5}]) + \mathbf{h}_S^{l-1}, \\ \mathbf{h}_{\text{FKA}}^{l} &= \text{MLP}(\mathbf{h}_{cr}), \end{aligned} \tag{2}$$

where $\mathbf{h}_{\text{FKA}}^{l-1}$ and $\mathbf{h}_T$ represent the output of the previous block of FKA and textual instructions, respectively. Self-A and Cross-A refers to the self and cross attention calculation in Transformer. Through the whole calculation of the FKA module, we can obtain the distilled visual knowledge $\mathbf{h}_{\text{FKA}}$ and will inject it into each block in LLMs to achieve in-depth interaction between LLMs and fine-grained visual knowledge.

As the right part shown in Figure 3, the input of the first layer in LLMs is $\mathbf{h}_{\text{IO}}, \mathbf{h}_{\text{KO}}$, and $\mathbf{h}_T$. We splice the sequence $\mathbf{h}_{\text{FKA}}$ according to the depth of LLMs and obtain the sequence of injected vectors: $\mathbf{h}_{\text{FKA}}^1, ..., \mathbf{h}_{\text{FKA}}^{LN}$, where *LN* represents the total number of layers for LLMs. The length of injected vectors is equal to N/LN and it will be spliced to the front of the input sequence. In summary, the whole framework is capable of visual knowledge alignment via VKA at the pretraining stage and the efficient injection of fine-grained visual knowledge via FKA.

3.4 Training

We pre-train the backbone of VKA and train VKA with the LLM on the same image-knowledge pairs from Wikipedia-based image text dataset WIT (Srinivasan et al., 2021) via the cross-entropy loss. Suppose that the training target is $Y = (y_1, ..., y_{\text{KM}})$, in which KM is the token length of knowledge description, the training objects are presented as the following two equations:

$$\begin{aligned} \mathcal{L}_{\text{VKA}}^{P} &= -\sum_{i=1}^{\text{KM}} \log P_i(\hat{y}_i = y_i | h_I, h_{\text{KP}}; \\ &\quad y_1, ..., y_{i-1}), \\ \mathcal{L}_{\text{VKA}}^{A} &= -\sum_{i=1}^{\text{KM}} \log P_i(\hat{y}_i = y_i | h_I, h_{\text{KP}}, h_{\text{KQ}}; \\ &\quad y_1, ..., y_{i-1}), \end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{VKA}}^{P}$ and $\mathcal{L}_{\text{VKA}}^{A}$ represents the loss of pre-training and aligning stages, respectively. While training CVLM on the multimodal instruction dataset, the overall training object is shown in Eq. 4

$$\mathcal{L}_{\text{CVLM}} = -\sum_{i=1}^{N_A} \log P_i(\hat{y}_i = y_i | I, T; t_1, ..., t_{i-1}), \tag{4}$$

where N_A and t_i refer to the total token count and the i th token of an answer.

4 Experiments

4.1 Data sets

Knowledge-based VQA is a task that requires reasoning with joint visual information, textual instructions, and outside knowledge. We mainly evaluate LMMs on the following relevant datasets. **OK-VQA (Marino et al., 2019)** is a visual question-answering dataset that requires methods that can draw upon outside knowledge to answer questions. It contains 9,009 training samples and 5,046 validation samples. **A-OKVQA (Schwenk et al., 2022)** is a knowledge-based visual multiple-choice question-answering benchmark that contains 17,056 training samples, 1,145 validation samples, and 6,702 testing samples. **VQAv2 (Goyal et al., 2017)** is a visual open-ended question-answering dataset where answering questions requires an integrated understanding of vision, language, and commonsense knowledge. It contains 443,757 training samples, 214,354 validation samples, and 447,793 test samples. **TextVQA (Singh et al., 2019)** is a task concerning reading and reasoning about text within images to answer questions related to them. The dataset comprises 34,602 training samples, 5,000 validation samples, and

Method	LLMs	Avg.	OK-VQA	VQAv2	A-OKVQA^{M}	A-OKVQA^{O}	TextVQA	InfoSeek	SEED-Bench^{S}
Promptcap (Hu et al., 2022)	GPT-3	-	-	73.2	56.3	-	-	-	-
Prophet (Yu et al., 2023)	mPLUG	-	-	76.6	64.7	-	-	-	-
VPD (55B) (Hu et al., 2023)	PaLI	-	-	84.7	62.7	-	-	-	-
Flamingo-9B	-	-	44.7	51.8	-	-	-	-	-
BLIP2	Flan-T5-XXL	-	45.9	65.2	-	53.71	-	10.67	-
MiniGPT4	Vicuna-7B	-	32.16	44.31	-	-	-	10.03	47.4
InstructBLIP	Flan-T5-XL	50.68	48.30	70.14	76.68	61.05	30.46	10.30	57.80
InstructBLIP	Flan-T5-XXL	50.83	46.91	70.11	78.34	62.01	29.79	8.36	60.29
InstructBLIP	Vicuna-7B	50.00	57.36	74.77	45.07	67.86	33.09	10.05	58.8
Qwen-VL (Bai et al., 2023)	Qwen	55.92	57.13	77.89	70.04	67.25	41.94	15.21	62.41
LLaVA-v1.5‡	Vicuna-7B	53.00	50.8	72.5	73.45	65.27	45.81	8.18	55.05
CVLM w/o (FKA & VKA)	Vicuna-7B	52.93	50.4	72.7	73.97	64.37	45.76	8.18	55.11
CVLM w/o FKA	Vicuna-7B	55.03	52.8	73.7	77.12	65.59	47.46	9.33	59.63
CVLM	Vicuna-7B	56.17	54.3	75.6	77.64	66.99	49.30	10.72	58.77
CVLM (3M IKPairs) w/o FKA	Vicuna-7B	57.19	55.7	75.7	77.90	70.22	49.89	11.21	59.73
CVLM (3M IKPairs, Objects=1)	Vicuna-7B	57.92	56.90	76.32	78.69	70.48	50.21	10.01	59.65
CVLM (3M IKPairs, Objects=3)	Vicuna-7B	58.19	57.17	76.40	79.21	70.91	50.32	10.42	62.93
CVLM (3M IKPairs Objects=5)	Vicuna-7B	57.83	56.92	76.48	78.95	70.39	50.48	11.27	60.30
CVLM (3M IKPairs, Objects=8)	Vicuna-7B	57.28	56.71	76.0	78.95	69.08	49.59	10.78	59.83
CVLM	Qwen-VL	**60.23**	58.91	80.88	82.71	72.14	45.02	15.45	66.28

Table 1: **Comparison between different LMMs on knowledge-based VQA benchmarks.** With 7B parameters, CVLM achieves the best performance with the same training data. "‡" shows that we fairly use the same instruction tuning data to train the model. "IKPairs" represents the image-knowledge pairs used to train VKA and the initial version is trained with 2M pairs. "Objects" refers to the number of object regions used in FKA, which are obtained by SAM. Benchmark names are abbreviated due to space limits. A-OKVQA^{M}: Multi-Choice A-OKVQA (Schwenk et al., 2022); A-OKVQA^{O}: Open-ended A-OKVQA (Schwenk et al., 2022); TextVQA (Singh et al., 2019); Infoseek (Chen et al., 2023b); SEED-Bench^{S}: SEED-Bench (Spatial) (Li et al., 2023b);

5,734 validation samples. *We integrated the training sets of the datasets above to construct a total of about 100K multi-turn instruction data*. Additionally, we also introduce the comprehensive evaluation benchmark SEED-Bench (Li et al., 2023a) and InfoSeek (Chen et al., 2023b) dataset to assess LMMs on comprehensive spatial understanding and fine-grained visual knowledge inferring.

4.2 Baselines

We mainly compare the proposed method to those current LMMs as follows: **BLIP2 (Li et al., 2023c)** is a Large Visual-Language Model that employs a Q-former to integrate visual features into the linguistic space. It provides detailed and accurate descriptions of visual content, effectively bridgiyng the gap between visual perception and linguistic understanding. **MiniGPT4 (Zhu et al., 2023)** utilizes the same pretrained vision components as BLIP-2, comprising a vision encoder and a Q-Former network. It introduces a single projection layer to align the encoded visual features with the Vicuna (Chiang et al., 2023). **InstructBLIP (Liu et al., 2023b)** utilizes the same Q-former to map visual information into the language space. Its specialization lies in understanding and responding to specific directives related to images, thereby enabling context-aware interactions with visual data. **LLaVA (Liu et al., 2023a)** connects the visual encoder with the Vicuna using a simple projection matrix. It could comprehend and generate multimodal content, seamlessly blending text and images for a holistic method of interpreting and generating varied forms of information.

4.3 Implementation Details

We utilize the AdamW (Kingma and Ba, 2014) optimizer with a cosine learning rate scheduler to train our model. During the pretraining stage of VKA, we first train it on 2 A100 GPUs using a dataset of **2 million image-knowledge pairs from Wikipedia** with a global batch size of 128 and a base learning rate of 5e-5. Image knowledge processing is shown in the *Appendix*. In the alignment stage, the model is trained on the same 2 million Wikipedia data using 2 A100 GPUs with a global batch size of 32 and a maximum learning rate of 1e-4. For the final stage, we employ LoRA to fine-tune the language model efficiently. In our implementation, we set the rank to 128 and alpha to 256, with a learning rate of 1e-4 for LoRA parameters and the newly added FKA. We use a smaller learning rate of 2e-5 for MLP and VKA.

Model	Avg.	Building	Animal	Plant	Location	Food	OC	Facility	Vehicle	Objects	Sport	Other
MiniGPT-4 (Vicuna-7b)	10.03	7.33	6.66	5.33	10.0	24.67	4.0	7.33	18.67	6.67	14.0	8.67
BLIP-2 (FlanT5-xxl)	10.67	8.7	2.67	4.0	16.0	14.0	9.33	16.0	28.0	2.0	9.33	7.33
InstructBLIP♣ (Vicuna-13b)	8.50	3.3	2.0	1.33	10.0	10.67	6.0	4.67	26.67	2.67	20.67	5.33
InstructBLIP♣ (FlanT5-xxl)	8.37	4.0	5.33	2.0	8.67	8.0	8.0	8.0	28.0	5.34	8.67	6.0
LLaVA-v1.5-13b♣	10.22	11.33	16.67	0.0	24.67	6.0	0.7	10.67	26.0	5.3	0.13	10.0
LLaVA-v1.5-7b‡♣	8.18	5.33	6.67	3.33	10.00	11.33	6.67	3.33	28.67	2.67	5.33	6.67
CVLM (LD=0)	9.33	3.33	14.67	5.33	6.0	14.0	6.0	2.67	36.67	4.0	0.67	9.33
CVLM (LD=2)	10.72	5.33	10.0	2.67	10.67	14.0	6.0	2.0	36.0	1.34	21.33	8.67
CVLM (LD=4)	9.94	4.0	8.0	2.0	9.33	15.33	4.67	2.67	38.0	1.33	16.67	7.33
CVLM (LD=8)	10.55	4.0	8.67	2.67	9.33	14.67	4.67	2.0	36.0	2.67	24.0	7.33
CVLM (3M IKPairs, LD=0)	11.21	4.67	10.0	5.33	8.67	15.33	5.44	3.33	38.0	3.33	22.67	6.67
CVLM (3M IKPairs, LD=2)	11.27	4.67	10.67	4.67	8.67	15.33	5.33	3.33	38.0	3.33	22.67	7.33
CVLM-624K (LD=0)	12.12	4.0	11.33	2.0	10.0	16.67	6.0	3.33	37.33	6.0	28.0	8.67
CVLM-624K (LD=2)	**12.30**	4.67	11.33	2.67	9.33	16.67	6.0	4.0	38.67	6.67	27.33	8.0

Table 2: **Held-out testing results on InfoSeek with fine-grained world knowledge**. Baseline results and knowledge categories are reported by Li et al. (2023i). "LD" represents the length of distillation vectors used in FKA. "LD=0" is identical to "w/o FKA". 'OC" refers to Organization and Company. ♣ indicates that the corresponding LMM baseline is trained using the training sets of knowledge-intensive datasets: OK-VQA and A-OKVQA.

Model	MMbench	ScienceQA-I	PoPE-Adversarial	PoPE-Random	PoPE-Popular
CVLM w/o (FKA & VKA)	50.85	62.76	82.94	86.98	86.14
CVLM w/o FKA	56.46	69.21	83.13	87.81	85.85
CVLM	**59.78**	68.92	82.78	88.79	85.64
CVLM (3M IKPairs) w/o FKA	57.06	69.96	82.32	85.75	84.26
CVLM (3M IKPairs)	**59.10**	69.85	82.47	86.90	84.68

Table 3: Comparison of different models across various common benchmarks. ScienceQA (Lu et al., 2022a) and PoPE (Li et al., 2023d) are used to evaluate the complex reasoning and hallucination recognition abilities of LMMs.

4.4 Main Results

Overall Performance. We present the comparative performance of all models across seven benchmarks tailored for knowledge-based VQA tasks, as detailed in Table 1. Our method exhibits a significant improvement over established baselines, with performance gains of 4% and 4.5% on dataset A-OKVQA and TextVQA, respectively. These advancements underscore the effectiveness of our visual-knowledge alignment modules VKA and FKA in bolstering the capabilities of LMMs, particularly evident in the enhancements to LLaVA-v1.5‡. Despite these improvements, our model slightly underperforms in comparison to InstructBLIP on the SEED-Bench, which may be attributed to the larger scale of multimodal instruction tuning data and larger language models (FlanT5-XXL-11B) used by InstructBLIP. We also evaluated our model on PoPE and ScienceQA in Table 3. On the ScienceQA dataset, we observe that including VKA results will largely improve the model performance yet FKA may not bring improvement. This may be attributed to the fine-grained features of images in ScienceQA containing less useful information. PoPE is an object recognition benchmark (only need the answer yes or no, like "Is there an apple in the image") and the similar performance indicates that introducing VKA and FKA will not affect the basic perception of an image.

Performance on Different Knowledge Categories. We evaluated the performance of our models across 11 detailed categories within the InfoSeek dataset, as outlined in Table 2. This fine-grained analysis reveals that our CVLM significantly outperforms existing models in specific categories, notably Animal, Vehicle, and Sport, showcasing its enhanced understanding and processing capabilities in these knowledge categories. Moreover, our comprehensive evaluation extends to the OK-VQA testing set, given in Table 4, further highlighting the impact of incorporating visual-knowledge alignment techniques. This strategic integration leads to notable improvements in knowledge-intensive VQA tasks, particularly in SR and PEL domains. All these results underscore the effectiveness of our approach in leveraging visual knowledge to enrich model performance across a spectrum of knowledge-driven categories.

Model	Avg.	VT	BCP	OMC	SR	CF	GHLC	PEL	PA	ST	WC	Other
MiniGPT-4 (Vicuna-7b)	29.31	28.67	31.03	26.0	28.0	25.33	38.21	22.67	29.33	29.23	31.25	34.0
BLIP-2 (FlanT5-xxl)	39.06	30.67	34.48	38.0	40.67	34.0	42.28	39.33	41.33	44.62	50.0	40.67
InstructBLIP♣ (Vicuna-13b)	41.02	34.00	52.41	37.33	51.33	33.33	46.34	31.33	38.67	32.30	49.11	43.33
InstructBLIP♣ (FlanT5-xxl)	47.96	44.66	51.03	48.67	48.0	43.33	51.22	47.33	42.0	55.38	58.04	45.33
LLaVA-v1.5-7b♣	57.25	50.0	62.76	58.0	62.67	54.0	60.16	50.0	53.33	61.54	65.18	57.33
LLaVA-v1.5‡♣	52.64	48.0	53.10	46.67	58.67	52.67	57.72	45.33	49.33	55.38	59.82	56.67
CVLM (LD=0)	55.92	49.33	62.07	53.33	61.33	49.33	62.60	47.33	50.67	60.0	68.75	57.33
CVLM (LD=2)	57.06	53.33	62.76	56.00	66.67	52.67	59.35	46.67	49.33	63.08	63.39	66.0
CVLM (LD=4)	56.25	50.67	61.38	53.33	60.0	52.67	58.54	47.33	50.0	63.08	64.29	64.0
CVLM (LD=8)	59.20	55.33	62.76	54.0	61.33	56.67	65.85	52.67	56.00	66.15	66.07	61.33
CVLM (3M IKPairs, LD=0)	58.79	59.33	65.52	50.67	63.33	54.67	65.04	52.67	50.67	61.54	65.18	62.67
CVLM (3M IKPairs, LD=2)	60.33	59.33	69.66	58.0	62.67	54.67	66.67	54.67	50.0	64.62	67.86	61.33
CVLM-624K (LD=0)	**61.47**	58.00	62.76	58.67	65.33	62.0	65.04	54.67	58.0	64.62	69.64	62.0
CVLM-624K (LD=2)	60.54	58.00	62.76	58.67	64.0	59.33	67.48	53.33	56.00	61.54	70.54	58.67

Table 4: **Held-In testing results on OK-VQA with Commonsense Knowledge**. Baseline results are reported by Li et al. (2023i). Knowledge names are abbreviated due to space limits. and Transportation (VT); Brands, Companies and Products (BCP); Objects, Material and Clothing (OMC); Sports and Recreation (SR); Cooking and Food (CF); Geography, History, Language and Culture (GHLC); People and Everyday Life (PEL); Plants and Animals (PA); Science and Technology (ST); Weather and Climate (WC); and Other.

4.5 Ablation Study

Effects of Visual Knowledge Aligner. To assess the impact of the Visual Knowledge Aligner on model performance, we trained LLaVA-v1.5 using the identical 504K dataset mentioned earlier. As depicted in Table 1, when compared to the baseline LLaVA-v1.5‡, CVLM(len=0) incorporating the Visual Knowledge Aligner yielded improved results across all benchmarks. Specifically, CVLM(len=0) exhibited the most significant enhancement on SEED-Bench(Spatial), achieving a 4.5% increase.

Effects of FKA. Then we study how the FKA influences the model performance. By comparing the experimental results of CVLM(len=0) and CVLM(len=2) on InfoSeek and OK-VQA benchmarks, and CVLM vs. CVLM w/o FKA in Table 1, we observe that the performance of the proposed method is further improved when FKA is added to the model. The reason for the improvement in our method's performance is that FKA enables the model to perceive fine-grained knowledge information, thereby further enhancing the understanding ability for knowledge-based questions.

Impact of the Size of Visual Knowledge Pairs. To validate the effectiveness of adding more knowledge pre-training data, we increased the amount of Wikipedia knowledge pre-training data from 2M to 3.3M during training VKA with LLM. The experimental results of CVLM (3M IKPairs) are shown in Tables 1, 2 and 4. We observed that the inclusion of more pretraining knowledge data significantly enhances the model's ability to comprehend knowledge, consequently resulting in higher performance on various tasks.

Analysis of Distillation Vector Length. Our examination of the optimal distillation vector length, as shown in Tables 2 and 4, indicates that increasing distillation vector length does not significantly improve model performance, rather results in performance fluctuations. This suggests that expanding distillation vectors could disturb the structural integrity of large language models, potentially reducing our model's effectiveness, especially in knowledge-dependent tasks.

Impact of introducing more instruction datasets. As the bottom results shown in Tables 2 and 4, we can see that more instruction data (624k from LLaVA-v1.5) will bring improvement on two knowledge-based VQA datasets, yet it will degrade the performance on some knowledge categories such as Plant. It indicates that introducing more data may not necessarily bring about an overall improvement in performance. Our cognitive mapper method leads to greater performance improvements with a small amount of instruction data, compared to adding more instruction data.

Analysis of the additional computation costs from SAM. We experimented with 1, 3, 5, and 8 image objects to determine the optimal count that maintains model performance while minimizing computational demand, as shown in Table 1. Our

Figure 4: Three cases illustrate the comparative performances of CVLM and other models. Red words represent the correct answer and the purple words show the inaccurate response.

findings suggest that extracting 3 or 5 regions offers a balanced approach. Increasing the number of image regions tends to introduce visual noise, which can either slightly degrade performance or offer negligible improvements. Furthermore, most evaluation samples require limited fine-grained visual information, featuring fewer objects.

4.6 Case Study

We present three cases in Figure 4 to thoroughly examine the performance of the models. Previous LMMs have been struggling with precise object identification in images and often providing generalized answers to knowledge-based questions. For example, as the case shown in the middle part of Figure 4, model InstructBLIP-vicuna-7B and LLaVA-7B incorrectly responded with "Wright Brothers" instead of naming the specific aircraft manufacturers shown. Case 3 demonstrated a similar limitation, with LLaVA-7B-LoRA merely predicting "Desert." However, employing the VKA and FKA mechanisms, CVLM displays superior capability in discerning crucial elements within images, thus furnishing more precise and contextually relevant responses based on pertinent knowledge.

5 Conclusion

In this work, we introduce the Cognitive Visual-Language Mapper (CVLM), an innovative approach that goes beyond the conventional alignment of visual and textual descriptions by incorporating visual-knowledge alignment. Specifically, we have developed a Visual Knowledge Aligner (VKA) that facilitates the projection of visual knowledge by acting as a bridge between the vision encoder and LLM. Additionally, we have integrated a Fine-grained Visual Knowledge Adapter (FKA) during the multimodal instruction tuning stage, which is designed to distill more precise knowledge pertinent to images and instructions. Our experimental findings demonstrate that CVLM outperforms several prominent LMMs that lack visual knowledge alignment. Our ablation studies highlight the effectiveness of VKA and FKA.

6 Acknowledge

Thanks for the efforts from reviewers and action editors. This work is supported by grants: Natural Science Foundation of China (No. 62376067).

Limitations

Our work has several limitations: 1) The knowledge representations gained by VKA may be inaccurate due to the loss of visual information and errors of knowledge association. Although we introduce large-scale visual knowledge data and the FKA to enhance visual knowledge acquisition, there is still potential to improve the accuracy of visual knowledge alignment. 2) From the experimental results, we observed that the distillation vector length impacts the stability of language models infused with visual knowledge information. Hence, we still need to explore an effective and stable visual knowledge-enhanced version of CVLM, especially for its FKA component. 3) The generated content may contain some factual errors or toxic statements due to the limitations of LLMs' generation capabilities. We also hope this work could spark further research on improving visual knowledge alignment during the construction of LMMs.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Patrick Cavanagh. 2011. Visual cognition. *Vision research*, 51(13):1538–1551.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. 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.

Cheng Chen, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, Yudong Zhu, and Xiaodong Gu. 2022. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18103–18112.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023a. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023b. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2023c. Temporal knowledge question answering via abstract reasoning induction. *arXiv preprint arXiv:2311.09149*.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023d. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392, Toronto, Canada. Association for Computational Linguistics.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jessica A. Collins and Ingrid R. Olson. 2014. Knowledge is power: How conceptual knowledge transforms visual cognition. *Psychonomic Bulletin and Review*, 21(4):843–860.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yushi* Hu, Hang* Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. *arXiv preprint arXiv:2312.03052*.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023e. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Yunxin Li, Baotian Hu, Xinyu Chen, Yuxin Ding, Lin Ma, and Min Zhang. 2023f. A multi-modal context reasoning approach for conditional inference on joint textual and visual clues. *arXiv preprint arXiv:2305.04530*.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023g. Lmeye: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701*.

Yunxin Li, Baotian Hu, Wei Wang, Xiaochun Cao, and Min Zhang. 2023h. Towards vision enhancing llms: Empowering multimodal knowledge storage and sharing in llms. *arXiv preprint arXiv:2311.15759*.

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, and Min Zhang. 2023i. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022b. Imagination-augmented natural language understanding. *NACCL*.

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

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. 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.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. *ICLR*.

OpenAI. 2023. Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*.

Steven Pinker. 1984. Visual cognition: An introduction. *Cognition*, 18(1-3):1–63.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

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

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

Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. 2023. Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. *arXiv preprint arXiv:2303.01903*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Shunyu Zhang, Xiaoze Jiang, Zequn Yang, Tao Wan, and Zengchang Qin. 2022. Reasoning with multi-structure commonsense knowledge in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4600–4609.

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

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.

A How Wikipedia knowledge pretraining data Constructed?

The construction of our Wikipedia knowledge pretraining data leverages images and their corresponding descriptions from approximately 2 million Wikipedia pages. Each page includes multiple images, enriched with Contextual Image Captioning and section summarization to serve as associated knowledge descriptions for the images. This comprehensive dataset is derived from work[1], providing a foundational resource for our pretraining data. Data Construction Processes are:

Data Processing: The initial step involves reading and parsing the TFRecord file. To manage the vast amount of data efficiently, concurrently, a thread pool is employed to download images based on the parsed image URLs.

Filtering: It was observed that approximately 1% of the image URLs were invalid. To ensure the quality and integrity of the dataset, these invalid entries were identified and subsequently removed from the dataset.

[1] WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning, 2021. https://github.com/google-research-datasets/wit/blob/main/wikiweb2m.md