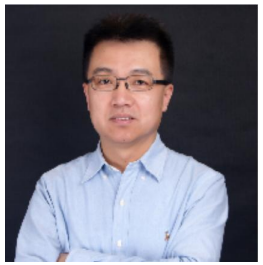
Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment

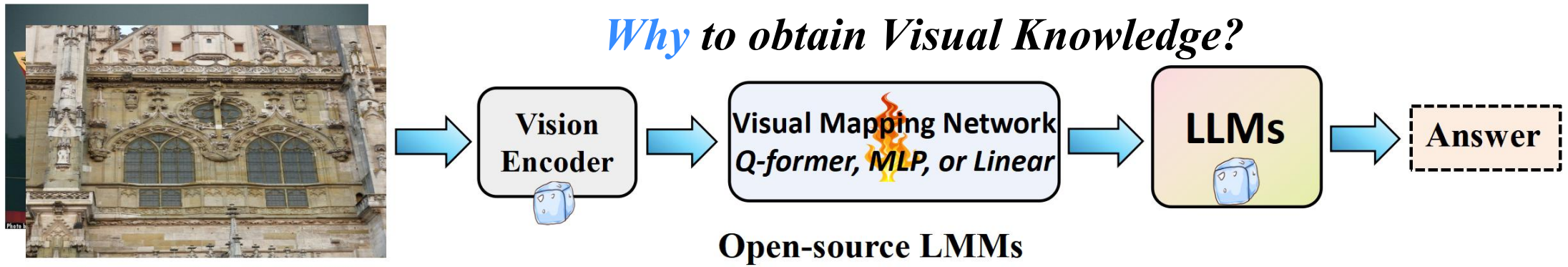
Yunxin Li, Xinyu Chen, Baotian Hu, Haoyuan Shi, Min Zhang



Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

Cognitive Visual-Language Mapper

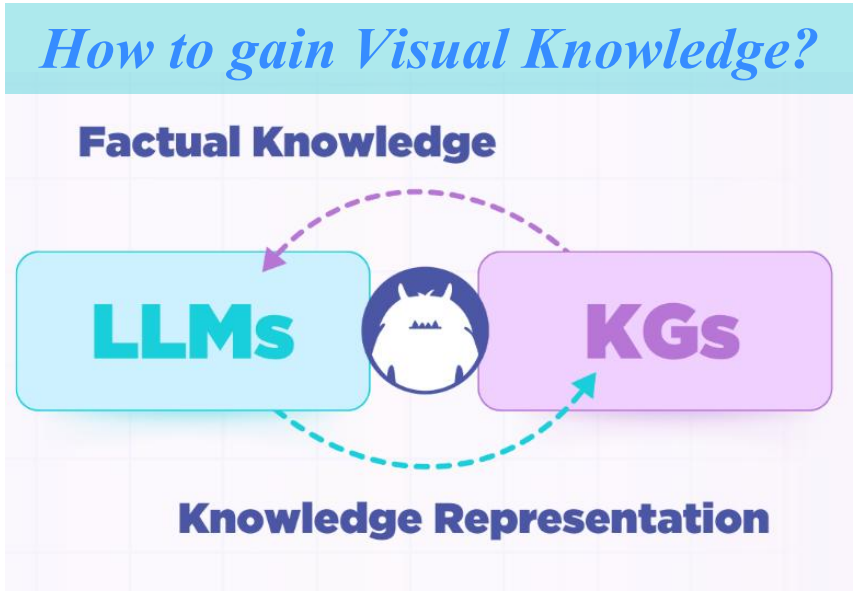
- Recent Large Multimodal Models (LMMs) such as GPT-4V and Gemini, have achieved impressive performance in a variety of visual understanding and reasoning tasks, especially on VQA.
- Current open-source LMMs are often constructed by **connecting visual encoders and large language models (LLMs) through a learnable visual mapping network**.
- Rethinking the construction process of LMMs, we find that
  - These visual mapping networks are trained on massive image-text captioning pairs and can only **transfer visual features to their corresponding language descriptions**.
  - They overlook linking an image with its relevant background knowledge, i.e., **connecting visuals to their knowledge.**

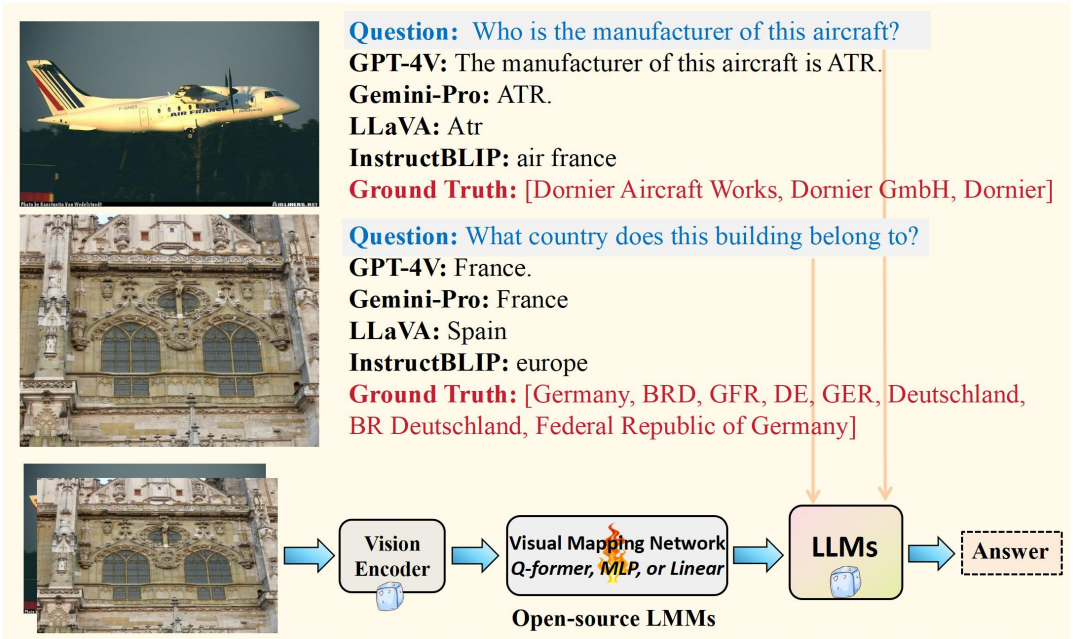


Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

Cognitive Visual-Language Mapper

- Visual knowledge plays a pivotal role in how 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.
- Previous works also demonstrated that introducing visual knowledge can improve the ability of LLMs.
- We present a **Cognitive Visual-Language Mapper (CVLM)**, which aims to explore improving LMMs with visual-language knowledge alignment





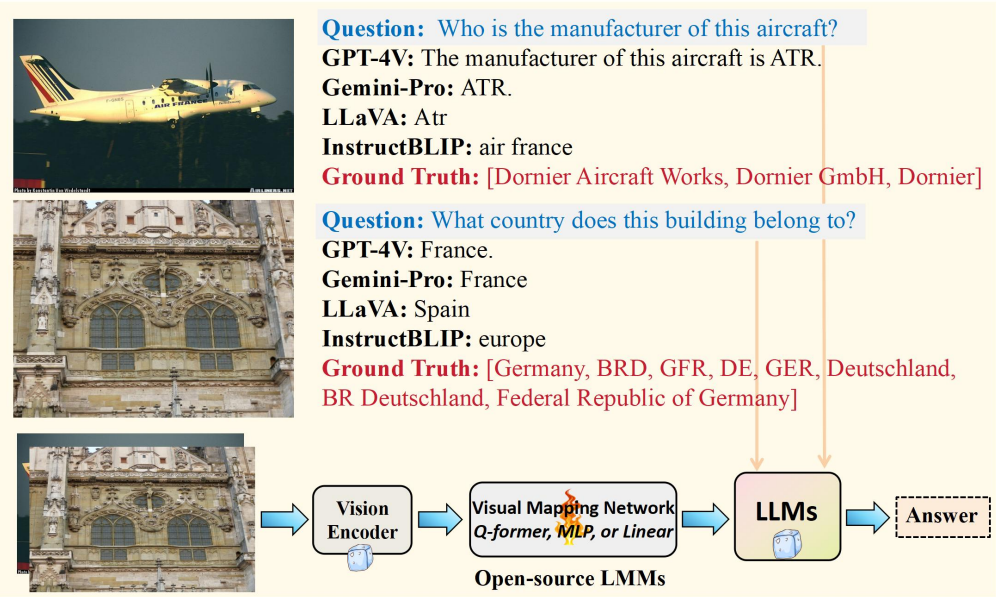
Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

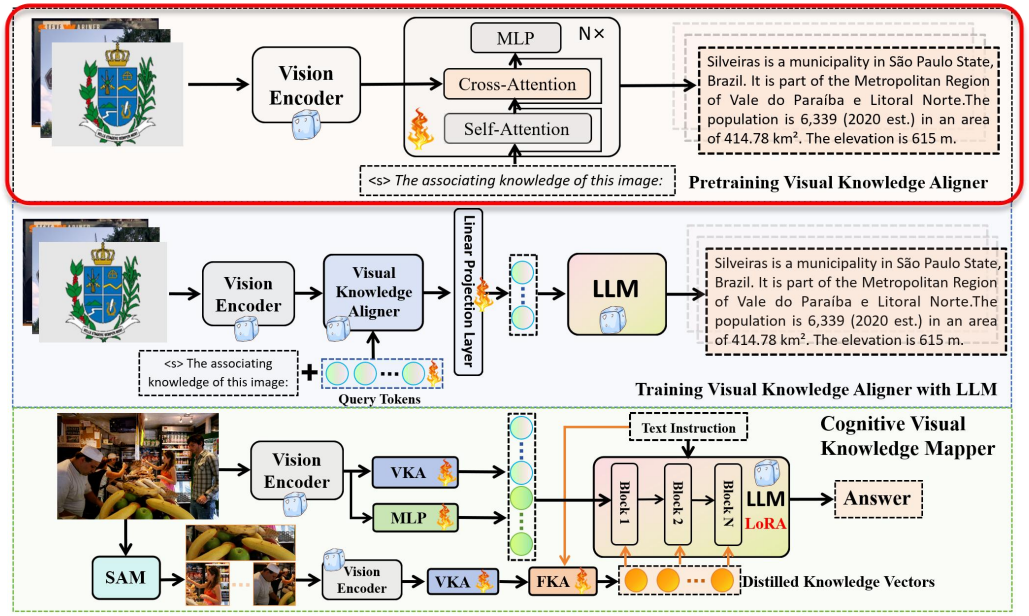
Cognitive Visual-Language Mapper

Architecture of CVLM

Stage 1: Pretraining Visual Knowledge Aligner (VKA):

1. A pretrained small language model (OPT) interacts with image representation by the cross-attention module.
2. It is trained to produce the relevant background knowledge of an image.
3. We train VKA with the image-knowledge pairs, which are collected from the Wikipedia pages.





Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

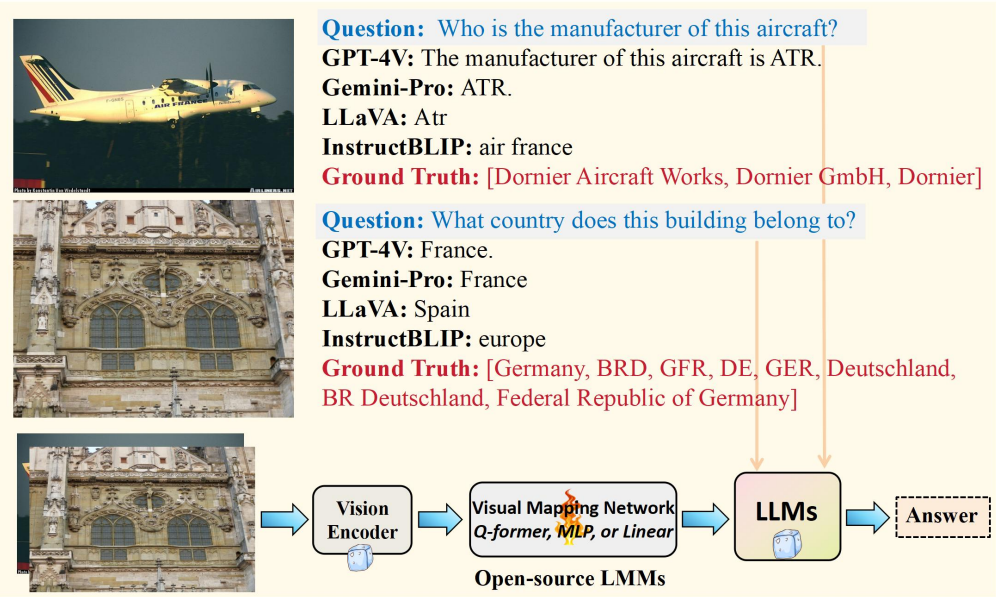
Cognitive Visual-Language Mapper

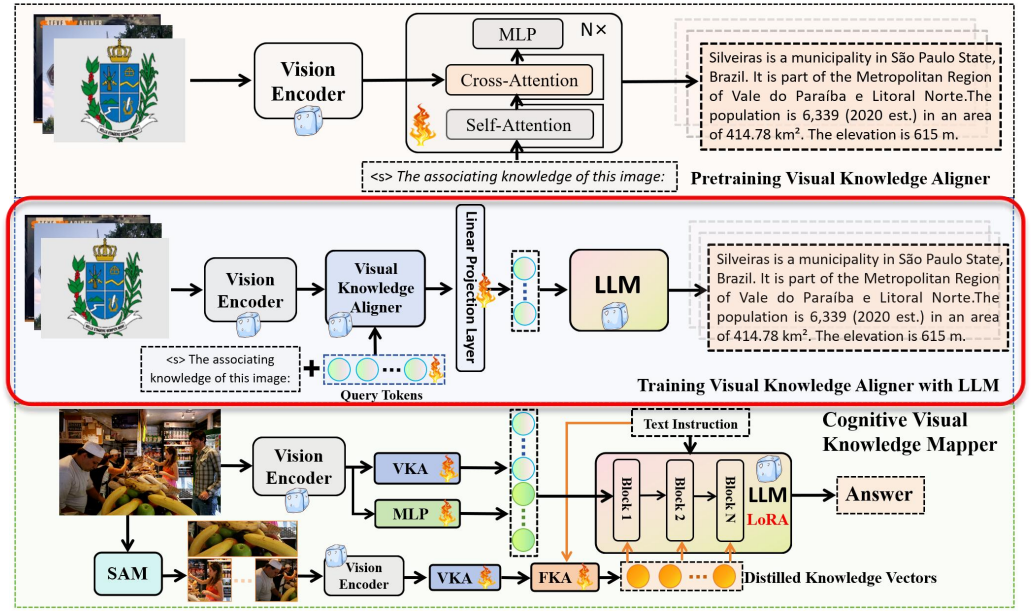
Architecture of CVLM

Stage 2: Training the VKA with Large Language Model:

1. We freeze the pretrained VKA, introduce a sequence of learnable query tokens and a linear projection layer to connect the VKA with LLM. We train it with the same visual-knowledge pairs as the previous pretraining stage.

2. This stage aims to make the LLMs receive representations that contain the knowledge of input images.





Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

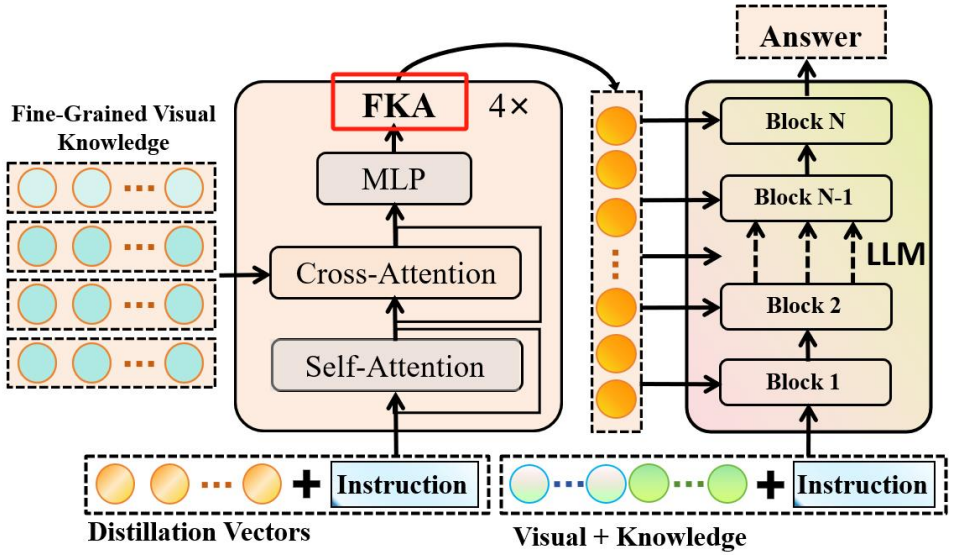
Cognitive Visual-Language Mapper

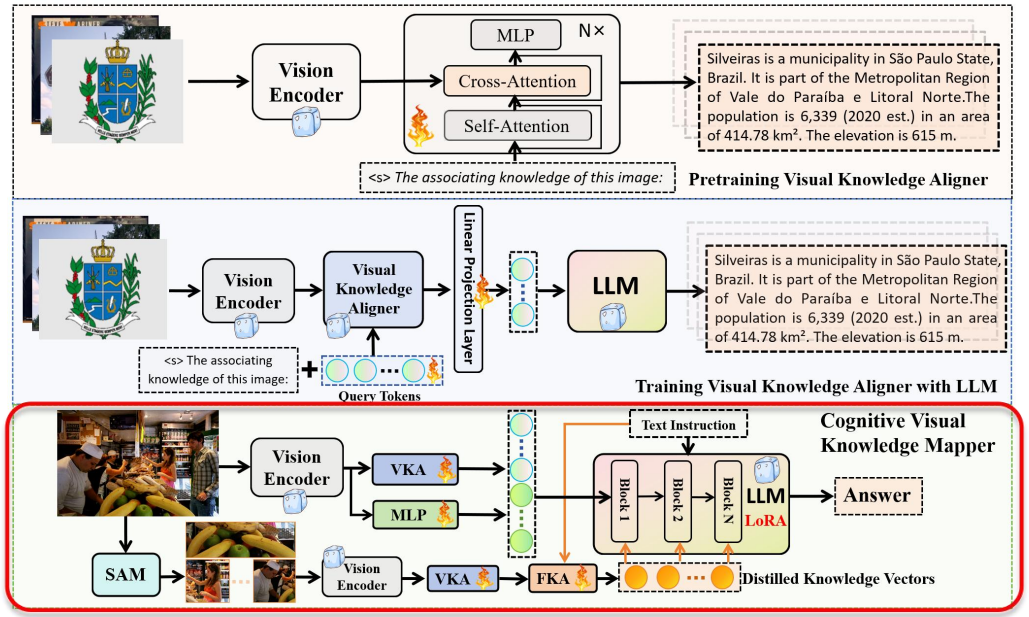
Architecture of CVLM

Stage 3: Fine-grained Knowledge Adapter (FKA):

1. We construct a fine-grained knowledge adapter based on the SAM (Image Segment) and the pretrained VKA.

2. It employs SAM to obtain multiple regions of an image and gain the relevant knowledge representation of these image regions by VKA. We distill these knowledge and inject them in each layer of LLMs, like Deep Prompting technical.





Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

Cognitive Visual-Language Mapper

Experimental Results

1. CVLM can adapt to the current mainstream LMMs and have an average improvement of **5 points on 7 VQA tasks**.
2. Fine-grained Knowledge Adapter can also improve the performance of LMMs, especially for **adopting 3 objects**.

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);

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.

Yunxin Li, et al. Cognitive Visual-Language Mapper: Advancing Multimodal Comprehension with Enhanced Visual Knowledge Alignment. *ACL 2024*.

Cognitive Visual-Language Mapper

Experimental Results

1. **More Visual Instruction data** can bring greater improvement for a MLLM with CVLM.
2. **Introducing More Visual-Knowledge pairs (2M→3M)** during the pretraining stage can lead to better performance.

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	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.

Cognitive Visual-Language Mapper

Case Analysis

Question: Where is this bird native to?
InstructBLIP-Vicuna-7B: south america
LLaVA-7B: Africa
LLaVA-7B -LoRA: South america.
CVLM: **Australia**
Knowledge Generated by VKA: *The Australian bird is a bird with a long beak and a long tail. It is a very popular bird in Australia, and is a popular symbol of Australian culture.*
Ground Truth: ['Australia', 'Aussieland', 'AU', 'Commonwealth of Australia']

Question: Who is the developer of this aircraft?
InstructBLIP-Vicuna-7B: wright
LLaVA-7B: Wright brother
LLaVA-7B -LoRA: Boeing
CVLM: **Supermarine**
Knowledge Generated by VKA: *The Supermarine Spitfire, a key British single-seat fighter aircraft of WWII, saw various versions from the Mk 1 to the Mk 22.*
Ground Truth: ['Supermarine', 'Vickers Supermarine']

Question: Where is this plant native to?
InstructBLIP-Vicuna-7B: mexico
LLaVA-7B: Africa
LLaVA-7B -LoRA: Desert
CVLM: **America**
Knowledge Generated by VKA: *mallow, is a flowering plant in the mallow family native to California.*
Ground Truth: ['the US', 'U.S', 'U.S.A.', 'United States of America']

Take Home Message

- Connecting Visuals to their relevant background knowledge can expand the understanding dimmension of visual information, improving the accuracy of answering information-seeking visual questions.
- Visual-Language Alignment technical still has a great improvement room towards understanding the objective world like a human.
- The current large multimodal model based on the large language model has a low correlation between the objective visual world and the high-level language knowledge summarized by humans, resulting in low accuracy in answering information-seeking questions.

Github: https://github.com/HITsz-TMG/Cognitive-Visual-Language-Mapper