Supplementary Material of LION: Empowering Multimodal Large Language Model with Dual-Level Visual Knowledge

Gongwei Chen, Leyang Shen, Rui Shao[†], Xiang Deng, Liqiang Nie[†]
School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
{chengongwei, shaorui, dengxiang, nieliqiang}@hit.edu.cn
https://github.com/rshaojimmy/JiuTian

A. Experimental Details

Architecture. We use the off-the-shelf ViT-G/14 from EVA-CLIP [2] without the last layer as our frozen vision backbone. The vision aggregator consists of two Bert Layers [1] with cross attention in each layer. The output from the Vision Aggregator undergoes a transformation via a two-layer MLP with GeLU [3] activation, and is projected into the latent feature space of the LLM. This output is then concatenated with the output from Q-Former and the textual inputs, forming the comprehensive inputs for the LLM. In the LLM, the hidden dimension of each adapter is set to 64. We implement LION on LLMs with two different size, including FlanT5-XL(3B) and FlanT5-XXL(11B), resulting in LION-4B and LION-12B, respectively.

When incorporating the image tags as high-level semantic visual evidence, we use the recognize anything model (RAM-14M) [5] based on the backbone Swin-Large. All the image tags are generated by using a 384×384 image size and a 0.8 threshold across 4585 categories in the ram tag list. All other hyperparameters are set the same as in the RAM codebase [*].

Training Details. Our training process comprises three stages. In Stage 1, we use a batch size of 64 for 10 epochs over 30k steps, with a learning rate starting at 1e-5 and reducing to a minimum of 0. Stage 2 increases the batch size to 256 for another 10 epochs across 60k steps, beginning with a learning rate of 5e-4, which is reduced to a floor of 1e-6; notably, the learning rate for the Vision Aggregator is set to a constant 1e-5. Stage 3 reverts to a batch size of 64 for 10 epochs and 60k steps, with an initial learning rate of 1e-5, descending to a minimum of 0. Throughout all stages, the AdamW [4] optimizer is employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.05. The learning rate is warmed up linearly from 1e-8 across 1000 steps at the beginning of each stage.

[†]Corresponding author
[*]https://github.com/xinyu1205/recognize-anything

Training Data. We describe all training datasets in Table 1. In stage 1, a part of LION is trained on image-level VL tasks, including COCO Caption, TextCaps, OKVQA, AOKVQA, VQAv2, OCR-VQA. Sepcifically, we follow InstructBLIP to define a visual question generation (VQG) task, which requires the model to generate a question given an answer. This VQG task is formed by using OKVQA, AOKVQA, and VQAv2 training datasets. We also use a dialogue dataset, LLaVA-Instruct-150K in this stage. In stage2, we use Visual Genome training dataset to construct referring expression comprehension (REC) and referring expression generation (REG) tasks. In final stage, all the mentioned datasets are used to train a unified model, resulting in the LION. We insert image tags in stage 3, firstly generate image tags for all training images, then use them with the soft prompting method. We provide evaluation metrics in Table 2.

B. Instruction Templates

B.1. Task Templates for Instruction-Tuning

We provide instruction templates for transform image-level and region-level VL tasks into a instruction-tuning format. For image-level VL tasks, we follow the setting in InstructBLIP. For region-level tasks, we use the templates in Shikra, which are generated by GPT-4 with carefully designed instructions. For each task listed in Table 3, we only show a few templates.

B.2. Instructions for Evaluation

We provide instructions for evaluation on various benchmarks. For instructions involving options, we arrange the options in the alphabetical order. For REC tasks, we randomly choose a template in training instruction lists for evaluation, which is the same as Shikra.

OKVQA, AOKVQA, GQA <Image> Question: {Question} Short answer:

COCOCap, Flickr30K, TextCaps <Image> A short image description:

Table 1. The training datasets used for instruction-tuning.

Task	Dataset	Stage 1	Stage 2	Stage 3	Data Number
Dialogue	LLaVA-Instruct-150K	✓		✓	361K
VQA	OKVQA, AOKVQA, VQAv2, OCR-VQA	✓		✓	1.3M
VQG	OKVQA, AOKVQA, VQAv2	✓		✓	470K
Image Captioning	COCO, TextCaps	✓		✓	524K
REC	Visual Genome		✓	✓	3.6M
REG	Visual Genome		✓	✓	3.6M

Table 2. Summary of the evaluation datasets.

Task	Dataset	Split	Metric
Image Captioning	Flickr30K	karpathy-test	CIDEr(↑)
	COCO	karpathy-test	CIDEr(↑)
	TextCaps	val	CIDEr(↑)
VQA	OKVQA	val	Accuracy(↑)
	AOKVQA	val	Accuracy(↑)
	Visual Spatial Reasoning	val	Accuracy(↑)
	Visual Dialog	val	MRR(↑)
	IconQA	test	Accuracy(↑)
	GQA	test-dev	Accuracy(↑)
REC	RefCOCO	val & testA & testB	Accuracy(↑)
	RefCOCO+	val & testA & testB	Accuracy(↑)
	RefCOCOg	val & test	Accuracy(↑)

IconQA <Image> {Question}
VSR <Image> Based on the image, is this statement true or false? "{Question}" Answer:
Visual Dialog <Image> Dialog history: {History}\n Question: {Question} Short answer:

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[2] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 1

[3] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1

[5] 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. 1

Table 3. Examples of instruction templates for various tasks. "{expr}" represents the expression in the REC task. "{BBox}" refers to the bounding box of a user-specified location.

VQA	<Image>Given the image, answer the following question with no more than three words. {Question} <Image>Based on the image, respond to this question with a short answer: {Question}. Answer: <Image>Use the provided image to answer the question: {Question} Provide your answer as short as possible: <Image>What is the answer to the following question? "{Question}" <Image>The question "{Question}" can be answered using the image. A short answer is
VQG	<Image>Based on the image, provide a question with the answer: {Answer}. Question: <Image>Given the visual representation, create a question for which the answer is "{Answer}". <Image>From the image provided, craft a question that leads to the reply: {Answer}. Question: <Image>Considering the picture, come up with a question where the answer is: {Answer}. <Image>Taking the image into account, generate an question that has the answer: {Answer}. Question:
Image Captioning	<Image>Can you briefly explain what you see in the image? <Image>Could you use a few words to describe what you perceive in the photo? <Image>Please provide a short depiction of the picture. <Image>Using language, provide a short account of the image. <Image>Use a few words to illustrate what is happening in the picture.
REC	<image>Identify the position of {expr} in image and share its coordinates. <image>I'd like to request the coordinates of {expr} within the photo. <image>How can I locate {expr} in the image? Please provide the coordinates. <image>I am interested in knowing the coordinates of {expr} in the picture. <image>Assist me in locating the position of {expr} in the photograph and its bounding box coordinates. <image>In the image, I need to find {expr} and know its coordinates. Can you please help?
REG	<image>What are the unique characteristics of the rectangular section {BBox} in image? <image>Describe the novel qualities of the selected bounding box {BBox} in image. <image>What sets the chosen region {BBox} in image apart from its surroundings? <image>Provide a one-of-a-kind depiction for the area enclosed by {BBox} in image. <image>How would you portray the unique features of the designated box {BBox} in image? <image>Explain the distinguishing characteristics of the marked bounding box {BBox} in image.