MoVE-KD: Knowledge Distillation for VLMs with Mixture of Visual Encoders

Supplementary Material

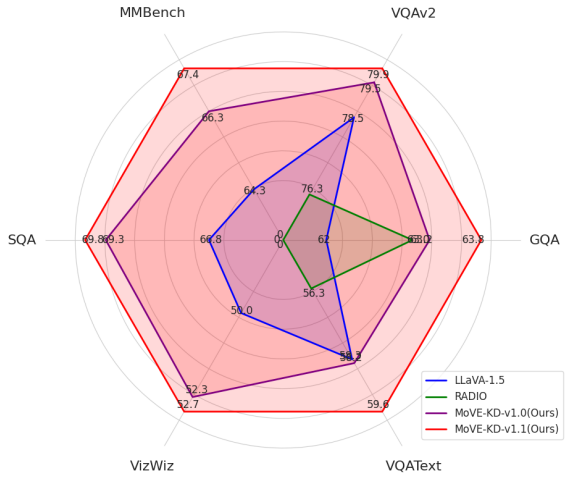


Figure 1. Comparison of MoVE-KD-v1.0 and MoVE-KD-v1.1.

A. Overview

In the appendix, we present more implementation details, additional experiments and visualization. For further insights, please refer to the anonymous website at https://MoVE-KD.github.io/, which includes a video demonstration.

B. Implementation Details

B.1. Dataset

We train the student model followed by the official dataset in LLaVA-1.5 [29] and LLaVA-NeXT [30], including VQA [14, 19, 36, 41], OCR [37, 45], region VQA [21, 23, 36], visual conversation [31] and language conversation [1].

B.2. Benchmarks

We conduct comprehensive experiments on several widely used visual understanding benchmarks, followed by a list.

VQA-v2. [14] The VQA-v2 benchmark evaluates the model's visual perception capabilities through open-ended questions. It consists of 265,016 images, covering various real-world scenes and objects, providing rich visual contexts for the questions. For each question, there are 10 ground truth answers provided by human annotators, which allows for a comprehensive evaluation of the performance of different models in answering the questions accurately.

GQA. [19] The GQA benchmark comprises three parts: scene graphs, questions, and images. The image part contains images, the spatial features of images, and the features of all objects in images. The questions in GQA are designed to test the understanding of visual scenes and the ability to reason about different aspects of an image.

TextVQA. [46] The TextVQA benchmark focuses on the comprehensive integration of diverse text information within images. It meticulously evaluates the model's text understanding and reasoning abilities through a series of visual question-answering tasks with rich textual information. Models need to not only understand the visual content of the images but also be able to read and reason about the text within the images to answer the questions accurately.

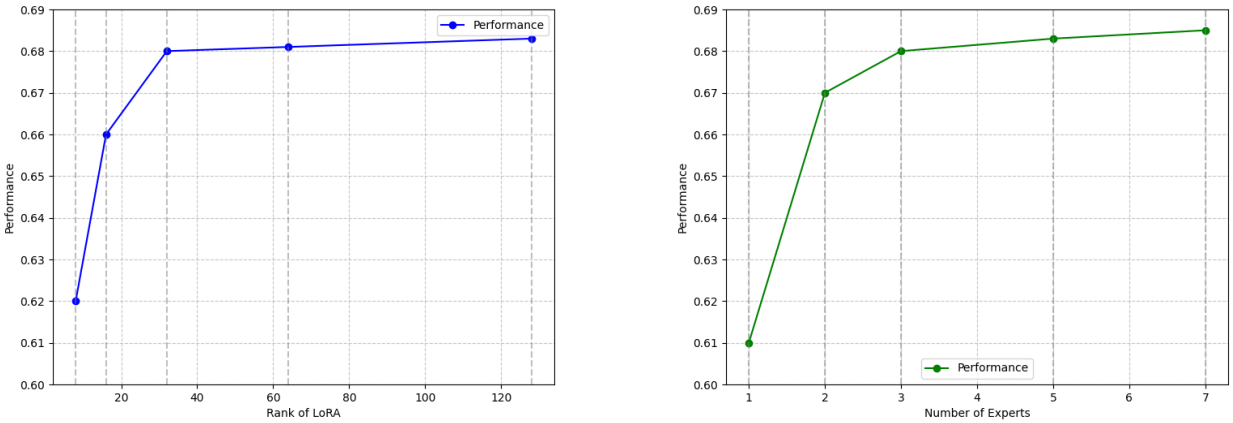
VizWiz. [15] The VizWiz dataset originates from a natural visual question-answering setting where blind people each took an image and recorded a spoken question, together with 10 crowd-sourced answers per visual question. The benchmark evaluates models to recognize, read, and reason about the visual elements and textual information in the images. This dual requirement challenges models to demonstrate robust text comprehension and logical reasoning skills, ensuring they can effectively integrate diverse information sources to provide accurate answers.

POPE. [24] The POPE benchmark is primarily used to evaluate the degree of Object Hallucination in models. It reformulates hallucination evaluation by requiring the model to answer a series of specific binary questions regarding the presence of objects in images. Accuracy, Recall, Precision, and F1 Score are effectively employed as reliable evaluation metrics to precisely measure the model's hallucination level under three different sampling strategies.

ScienceQA. [34] The ScienceQA benchmark covers diverse domains, including natural science, language science, and social science. Within each subject, questions are categorized first by the topic, then by the category, and finally by the skill. This hierarchical categorization results in 26 topics, 127 categories, and 379 skills, providing a comprehensive and diverse range of scientific questions. It provides a comprehensive evaluation of a model's capabilities in multimodal understanding, multi-step reasoning, and interoperability.

MME. [12] The MME benchmark is also a comprehensive benchmark meticulously designed to thoroughly evaluate various aspects of a model's performance. It consists of 14 sub-tasks that specifically aim to evaluate both the model's perceptual and cognitive abilities. By utilizing manually constructed instruction-answer pairs and concise instruction design, it effectively mitigates issues such as data leakage and unfair evaluation of model performance.

MMBench. [32] The MMBench benchmark comprehensively evaluates the model's overall performance across



(a) Impact of the rank of LoRA

(b) Impact of number of experts

Figure 2. The impact of the rank of LoRA and the number of experts on MoLE performance. Marginal benefits diminish when the LoRA rank exceeds 32 or the number of experts is greater than 3.

multiple dimensions by incorporating multiple-choice questions. It includes three levels of ability dimensions. The first level (L1) consists of two main abilities, perception and reasoning. The second level (L2) expands based on the first level, including six sub-abilities. The third level (L3) further refines the second level, encompassing 20 specific ability dimensions. This hierarchical structure enables a comprehensive evaluation of the model's various capabilities.

B.3. Configurations

- **Input resolution:**
  - **CLIP:** 336×336
  - **EVA:** 1024×1024
  - **ConvNeXT:** 1024×1024
  - **SAM:** 1024×1024
- **Checkpoints:**
  - **CLIP:** clip-vit-large-patch14-336
  - **EVA:** eva02_L_coco_det_sys_o365.pth
  - **ConvNeXT:** CLIP-convnext_xxlarge-laion2B-s34B-b82K-augreg-soup
  - **SAM:** sam-vit-large
- **MoLE:**
  - **Number of experts:** 3
  - **Rank of LoRA:** 32

C. Additional experiments

C.1. Better performance with more teachers

To further demonstrate the scalability of MoVE-KD and the impact of additional teachers on the performance, we employ the SAM-L [22] as a new teacher alongside the original ones. We denote the two versions as MoVE-KD-v1.0 and MoVE-KD-v1.1. As shown in Table 1, MoVE-KD-v1.1 further improves performance and outperforms MoVE-KD-v1.0 on VQA^{Text}, with MoVE-KD-v1.1 7B even surpassing LLaVA-1.5 13B on some benchmarks. However, as the size of the LLM increases, we observe diminishing performance gains from vision encoder distillation, which aligns with our previous conclusions. Therefore, MoVE-KD demonstrates strong compatibility with advanced encoders, and in the future, we will continue to explore a teacher pool of visual encoders more suitable for VLMs.

C.2. Comparison of MSE-based distillation

In the ablation study, we demonstrate the gap between the distillation method based on simple interpolation with mean squared error (MSE) loss and MoVE-KD. Due to space constraints in the main text, we present this method (referred to as MSE) as a baseline in Tab. 2 to highlight the superiority and effectiveness of our proposed improvements.

C.3. Settings of MoLE

We further investigated the impact of the rank of LoRA and the number of experts in MoLE on the experimental results, using the average performance across various benchmarks (excluding MME [12]) as the evaluation metric. As shown in Fig. 2, when the rank of LoRA exceeds 32 and the number of experts exceeds 3, the marginal benefits diminish. Therefore, to balance performance and computational cost, we choose the most appropriate hyperparameter settings.

D. Visualization

In Fig. 3, we additionally show the results without [CLS] attention regularization, which shows that the [CLS] attention has a low correlation with key information.

Method	LLM	VQA^{V2}	GQA	VQA^{Text}	VizWiz	POPE	SQA	MME	MMB
1.7B Models									
LLaVA-1.5 [29]	MobileLLaMA-1.4B [9]	71.5	55.4	42.6	28.6	84.3	56.0	1145.7	47.0
+ MoVE-KD-v1.0	MobileLLaMA-1.4B [9]	72.9	56.6	43.4	**32.1**	84.8	56.1	1182.3	47.4
+ MoVE-KD-v1.1	MobileLLaMA-1.4B [9]	**73.8**	**57.7**	**44.3**	29.3	**86.1**	**57.3**	**1188.4**	**48.8**
7B Models									
LLaVA-1.5 [29]	Vicuna-7B [59]	78.5	62.0	58.2	50.0	85.9	66.8	1510.7	64.3
+ MoVE-KD-v1.0	Vicuna-7B [59]	79.5	63.2	58.3	52.3	**86.9**	69.3	**1524.5**	66.3
+ MoVE-KD-v1.1	Vicuna-7B [59]	**79.9**	**63.9**	**59.6**	**52.7**	86.3	**69.8**	1509.1	**67.4**
13B Models									
LLaVA-1.5 [29]	Vicuna-13B [59]	80.0	63.3	61.3	53.6	85.9	71.6	1531.3	67.7
+ MoVE-KD-v1.0	Vicuna-13B [59]	80.6	**64.2**	59.7	55.7	85.7	**73.2**	1568.1	**70.2**
+ MoVE-KD-v1.1	Vicuna-13B [59]	**80.8**	63.9	61.1	**57.5**	**86.3**	71.8	**1568.3**	69.7

Table 1. **Performance of MoVE-KD-v1.0 and MoVE-KD-v1.1**

Method	LLM	VQA^{V2}	GQA	VQA^{Text}	VizWiz	POPE	SQA	MME	MMB
1.7B Models									
LLaVA-1.5 [29]	MobileLLaMA-1.4B [9]	71.5	55.4	42.6	28.6	84.3	56.0	1145.7	47.0
+ MSE	MobileLLaMA-1.4B [9]	72.0	55.8	42.5	29.3	84.2	55.4	1161.8	47.1
+ MoVE-KD (Ours)	MobileLLaMA-1.4B [9]	**72.9**	**56.6**	**43.4**	**32.1**	**84.8**	**56.1**	**1182.3**	**47.4**
7B Models									
InstructBLIP [31]	Vicuna-7B [59]	-	49.2	50.1	34.5	-	60.5	-	36
Qwen-VL [5]	Qwen-7B [5]	78.8	59.3	63.8	35.2	-	67.1	1487.5	38.2
LLaVA-1.5 [29]	Vicuna-7B [59]	78.5	62.0	58.2	50.0	85.9	66.8	1510.7	64.3
+ MSE	Vicuna-7B [59]	79.0	62.4	56.7	50.9	84.7	67.6	1507.6	62.9
+ RADIO [40]	Vicuna-7B [59]	76.3	63.0	56.3	-	86.2	-	-	-
+ MoVE-KD (Ours)	Vicuna-7B [59]	**79.5**	**63.2**	**58.3**	**52.3**	**86.9**	**69.3**	**1524.5**	**66.3**
LLaVA-NeXT [30]	Vicuna-7B [59]	81.8	64.2	64.9	57.6	86.5	70.1	1519.0	67.4
+ MSE	Vicuna-7B [59]	82.0	63.5	63.1	57.2	86.4	70.3	1526.9	67.0
+ MoVE-KD (Ours)	Vicuna-7B [59]	**82.3**	**64.5**	63.7	**58.0**	**86.7**	**70.7**	**1537.2**	**67.6**
13B Models									
InstructBLIP [31]	Vicuna-13B [59]	-	49.5	50.7	33.4	78.9	63.1	1212.8	-
LLaVA-1.5 [29]	Vicuna-13B [59]	80.0	63.3	61.3	53.6	85.9	71.6	1531.3	67.7
+ MSE	Vicuna-13B [59]	80.1	63.7	58.4	54.6	84.7	71.9	1543.3	68.8
+ MoVE-KD (Ours)	Vicuna-13B [59]	**80.6**	**64.2**	59.7	**55.7**	85.7	**73.2**	**1568.1**	**70.2**
LLaVA-NeXT [30]	Vicuna-13B [59]	82.8	65.4	67.1	60.5	86.2	73.6	1575.0	70
+ MSE	Vicuna-13B [59]	82.4	65.1	64.5	60.7	86.3	73.4	1568.4	70.3
+ MoVE-KD (Ours)	Vicuna-13B [59]	**83.1**	**65.7**	65.8	**60.9**	**86.8**	**73.7**	**1579.3**	**70.6**

Table 2. **Performance of MoVE-KD and other methods** (inlcude MSE KD.)

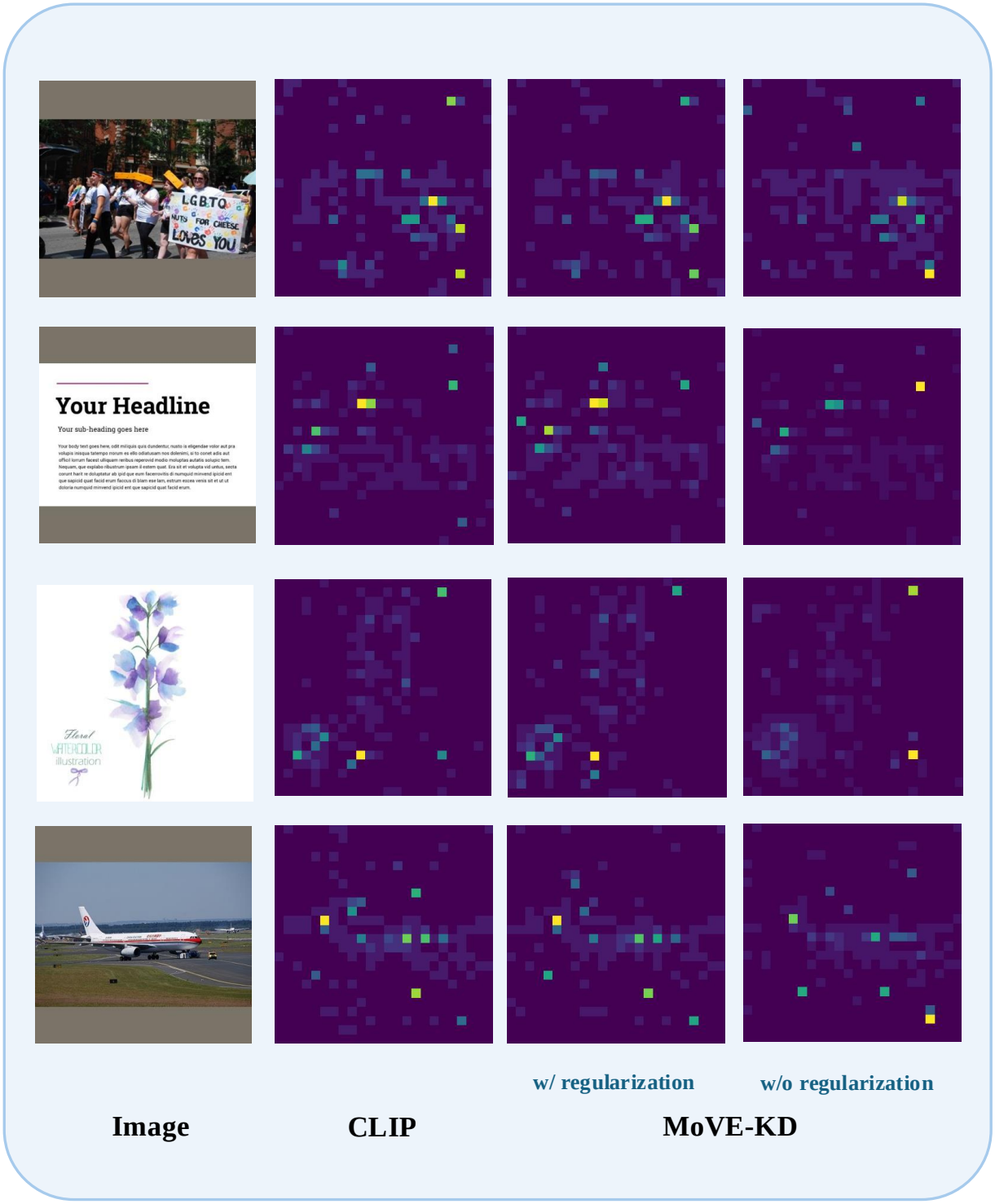


Figure 3. **The visualization of the [CLS] attention of CLIP and MoVE-KD (w/ and w/o regularization).**