LLaVA-KD: A Framework of Distilling Multimodal Large Language Models

Yuxuan Cai[1*] Jiangning Zhang[2,3*] Haoyang He[2] Xinwei He[4] Ao Tong[1] Zhenye Gan[3]
Chengjie Wang[3] Zhucun Xue[2] Yong Liu[2] Xiang Bai[1†]
[1]Huazhong University of Science and Technology [2]Zhejiang University
[3]Tencent Youtu Lab [4]Huazhong Agricultural University
{cyx_hust, xbai}@hust.edu.cn, 186368@zju.edu.cn

Supplementary Material Overview

The supplementary material presents more comprehensive analysis and results of our LLaVA-KD:

- **Sec.** A.1 presents a comparative analysis between LLaVA-KD and the state-of-the-art MLLM distillation approach LLaVA-MoD.
- **Sec.** A.2 analyzes the critical role of visual token representations during the distillation process.
- **Sec.** A.3 provides comprehensive comparisons between LLaVA-KD and existing small-scale MLLMs in terms of training data, computational efficiency, and performance.
- **Sec.** A.4 investigates the model's robustness under various hyperparameter.
- **Sec.** A.5 provides qualitative comparisons between LLaVA-KD and TinyLLaVA baseline.
- **Sec.** B.1 provides more detailed results of ablation study on our proposed distillation strategies: Multimodal Distillation (MDist) and Relation Distillation (RDist).
- **Sec.** B.2 provides more detailed results of ablation study on the distillation targets.
- **Sec.** B.3 provides more detailed results of the validation on another MLLM framework.
- **Sec.** B.4 provides more detailed results of ablation study on teacher models with different sizes.
- **Sec.** C provides the implementation details.

A. More Ablation and Explanatory Analysis

A.1. Comparison and discussion with LLaVA-MOD

LLaVA-MOD [6] represents a state-of-the-art approach in MLLM distillation. As shown in Table A1, we evaluate our method using the same benchmarks as LLaVA-MOD. We can observe that LLaVA-KD demonstrates significant performance improvements at equivalent model scales. Specifically, it achieves average gains of 1.4% and 1.5% for 1B and 2B student models, respectively. It is important to note that LLaVA-MoD's training data is nearly 3.8 million more than ours (5M vs. 1.2M). Despite this disparity, our method still outperforms, which further emphasizes our method's efficiency and effectiveness.

We compare our approach with LLaVA-MoD [6] to highlight the technical differences:

- **Architecture Design.** While LLaVA-MoD employs Mixture-of-Experts (MoE) [4] layers for its student model, we maintain a simple yet effective architecture for s-MLLM without additional complexity.
- **Training Scheme.** We propose a three-stage training framework: *1)* Distilled Pre-Training (DPT) to promote the visual-textual alignmenet, *2)* Supervised Fine-Tuning (SFT) for knowledge acquisition, *3)* and the Distilled Fine-Tuning (DFT) to transfer l-MLLM's knowledge to the s-MLLM. In contrast, LLaVA-MoD first follows the conventional Pre-Training, followed by two-stage distillation pipeline: Mimic Distillation for knowledge transfer and Preference Distillation for hallucination reduction.
- **Distillation Strategy.** We introduce dedicated knowledge distillation strategies (MDist/RDist) for both DPT and DFT stages, whereas LLaVA-MoD introduces the Preference Optimization (PO) loss [2] during the Preference Distillation stage.

A.2. Further discussion on visual tokens

As demonstrated in Table A2, we investigate the effects of visual token distillation in our LLaVA-KD framework. This process involves two key aspects. First, we employ the visual token matching loss (L_{vis}) to align the student model (s-MLLM) with the teacher model (l-MLLM) through Kullback-Leibler divergence minimization on their visual token distributions. Second, the relational distillation loss (L_{rel}) which matches the inter-token relationship matrices between the s-MLLM and l-MLLM, enhancing the visual token quality. Our experiments demonstrate that visual token distillation yields consistent performance gains: +0.7% average improvement during the DPT+SFT phase, with an additional +0.7% enhancement in the DFT phase.

*Equal Contribution. Work done during internship at Tencent YouTu Lab. †Correspondence to Xiang Bai<xbai@hust.edu.cn>.

Method	LLM of Teacher	LLM of Student	Image Question Answering				Benchmarks			Avg_7
			GQA	VizWiz	SQA	TextVQA	MME	MMB	MMB^{CN}	
LLaVA-MoD	Qwen1.5-4B	Qwen1.5-1.8B	58.7	34.6	67.9	57.7	67.6	64.9	60.7	58.9
LLaVA-KD			62.3	44.7	64.7	53.4	69.1	64	63.7	60.3
LLaVA-MoD		Qwen1.5-0.5B	56.0	25.3	64.7	53.8	63.3	62.2	50.8	53.7
LLaVA-KD			59.6	35.9	60.6	49.9	64.5	60.1	55.5	55.2

Table A1. Comparison with LLaVA-MoD. Avg_7: The average performance of the seven benchmarks.

Training Stage	Distillation Loss	Image Question Answering					Benchmarks					Avg_{10}
		VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
DPT-SFT	L_{res}	73.8	57.8	25.6	**62.8**	47.1	59.7	55.9	49.3	85.5	**31.6**	54.9
DPT-SFT	$L_{res} + L_{vis} + L_{rel}$	74.6	57.8	28.6	51.2	49.1	59.9	56.9	51.6	84.3	31.4	55.6
DPT-SFT-DFT	L_{res}	76.8	**59.6**	**36.4**	59.1	**50.4**	57.6	52.7	54.2	85.7	30.1	57.2
DPT-SFT-DFT	$L_{res} + L_{vis} + L_{rel}$	**77.0**	**59.6**	35.9	60.6	49.9	**64.5**	**60.1**	**55.5**	**85.9**	30.2	**57.9**

Table A2. Influence of distilling visual tokens during Model training process. Avg_{10}: The average performance of the ten benchmarks. Optimal and sub-optimal results are in **bold** and underline, respectively.

Method	#Params	#Samples	Time	Avg_7	Avg_7 (+POPE)	Avg_7 (+POPE+VQAv2)	Avg_7 (+POPE+VQAv2+MMMU)
TinyLLaVA	∼2B	1.2 M	105	53.9	57.6	59.3	56.8
MoE-LLaVA		2.2 M	/	55.3	59.2	61.1	-
Bunny		2.6M	/	56.3	60.0	61.8	-
Mini-Gemini		2.7M	/	57.1	60.7	-	-
Imp		1.5M	/	58.9	62.4	64.3	-
LLaVA-MoD		5 M	960	59.9	63.3	-	-
LLaVA-KD		1.2 M	320	**60.3**	**63.5**	**65.2**	**62.1**
TinyLLaVA	∼1B	1.2 M	52	51.1	55.2	57.3	54.7
SPHINX-Tiny		15 M	/	51.2	55.1	57.3	-
LLaVA-MoD		5 M	/	54.1	-	-	-
LLaVA-KD		1.2 M	210	**55.2**	**59.0**	**61.0**	**57.9**

Table A3. Efficiency comparison of SoTA MLLMs. The "Avg_7" is calculated on seven benchmarks (excluding POPE, VQAv2, and MMMU), while subsequent columns incrementally incorporate POPE, VQAv2, and MMMU to evaluate comprehensive capabilities.

α, β, γ	α', β', γ'	Image Question Answering					Benchmarks					Avg_{10}
		VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
{1,1,0.5}	{1,1,0.5}	77.0	59.6	39.5	60.6	49.9	64.5	60.1	55.5	85.9	39.2	57.9
{1,1,0.5}	{1,0.5,0.5}	76.7	59.5	36.6	59.4	49.3	63.9	58.2	55.1	84.5	39.3	57.8
{1,1,5}	{1,1,5}	76.8	59.8	35.7	60.4	50.9	62.1	59.4	54.7	86.0	31.7	57.8

Table A4. Influence of hyperparameters during model training process. Avg_{10}: The average performance of the ten benchmarks.

This progressive improvement underscores the effectiveness of visual token distillation.

A.3. Comparison with SoTA small MLLMs

In Table A3, we compare our model with SoTA small-scale MLLMs in terms of model size (#Params), training samples (#Samples) and training time (Time). The "Avg_7" is computed on seven benchmarks (excluding POPE, VQAv2 and MMMU) to enable direct model comparisons.

With 1B parameters, LLaVA-KD achieves superior performance with 4.0% and 1.1% improvements over SPHINX-Tiny [3] and LLaVA-MoD [6] respectively, despite requiring fewer training samples. This efficiency advantage persists in 2B-parameter models, where our method outperforms the previous art Imp [5] and LLaVA-MoD by margins of 1.4% and 0.4%. In addition, compared to our baseline TinyLLaVA [7], despite an increase in training time, LLaVA-KD delivers substantial performance gains of 4.1% (1B) and 6.4% (2B), respectively.

We further incrementally incorporate the three benchmarks for comprehensive comparison (Cols 6-8). It can be observed that our method demonstrates consistent superiority among models of the same scale. Overall, our experimental results demonstrate that LLaVA-KD achieves an optimal balance between training efficiency and model performance compared to existing s-MLLM models.

A.4. Analysis of Hyperparameters

To verify the robustness of our method, we analyze the sensitivity of model performance to hyperparameter configurations. The key hyperparameters consist of loss weighting coefficients (α, β, and γ) in the DPT stage and (α', β' and γ') in the DFT stage, which balance the contributions of L_{res}, L_{vis} and L_{rel} within their respective training ob-

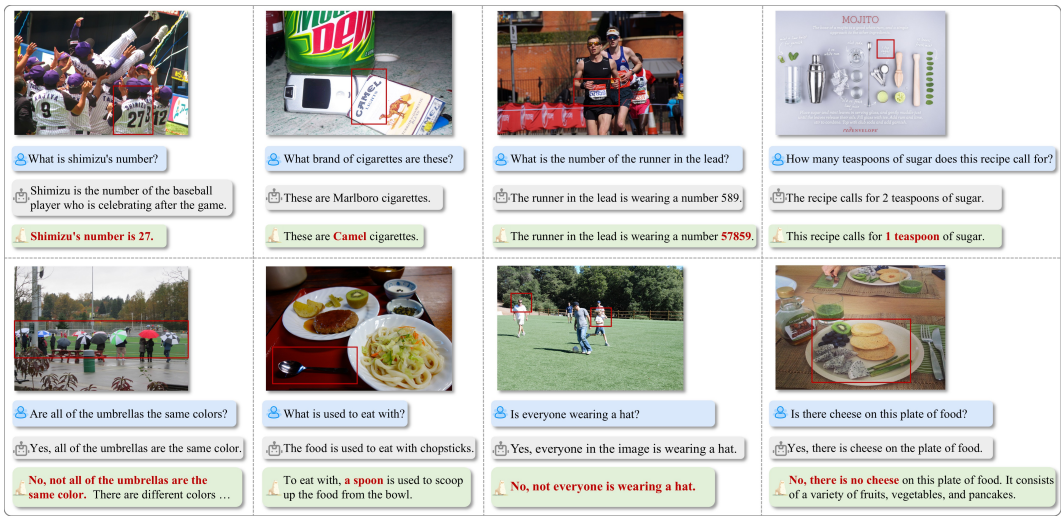

Figure A1. Qualitative visualization comparison between our LLaVA-KD with TinyLLaVA.

Distilled Pre-Training		Image Question Answering					Benchmarks					Avg_{10}
MultiModal Distill	Relation Distill	VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
✗	✓	73.6	53.3	39.7	59.0	47.6	54.4	58.5	55.0	84.4	30.0	55.5
✓	✗	74.5	58.3	26.7	62.6	48.5	57.3	57.1	48.6	85.6	31.8	55.1
✓	✓	74.6	57.8	28.8	61.2	49.1	59.9	56.9	51.6	84.3	31.4	55.6

(a) Different distillation strategies during the DPT stage.

Distilled Fine-Tuning		Image Question Answering					Benchmarks					Avg_{10}
MultiModal Distill	Relation Distill	VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
✗	✓	76.1	58.6	37.4	59.7	49.1	60.6	58.5	53.9	86.2	30.0	57.0
✓	✗	76.9	59.7	38.3	59.9	49.4	64.1	57.4	54.8	86.3	30.7	57.7
✓	✓	77.0	59.6	35.9	60.6	49.9	64.5	60.1	55.5	85.9	30.2	57.9

(b) Different distillation strategies during the DFT stage.

Table A5. Ablation study on Multimodal Distillation and Relation Distillation during DPT and DFT stages.

jectives L_{DPT} and L_{DFT}. We conduct an ablation study with several hyperparameter combinations. As shown in Table A4, our model exhibits marginal variations in average performance across different configurations. The results confirm the robustness of our LLaVA-KD to hyperparameter choices.

A.5. Visualization

Fig. A1 shows qualitative results between our LLaVA-KD-1B and the baseline TinyLLaVA-1B [7]. It can be observed that our approach achieves a more accurate understanding of multimodal information, leading to more precise responses.

B. Detailed Results

B.1. Ablation study on Distillation strategy

Table A5 shows more detailed results of the ablation study on the distillation strategy, including MDist and RDist. Table A5(a) evaluates performance after the DPT-SFT training phase, while Table A5(b) conducts ablation studies at the DFT stage.

B.2. Aablation study on Distillation targets

Table A6 shows more detailed results of the ablation study on the distillation targets in our MDist distillation strategy. Table A6(a) evaluates performance after the DPT-SFT training phase, while Table A6(b) conducts ablation studies at the DFT stage.

B.3. Validation on MobileVLM

Table A7 shows more detailed results of the ablation study that employs our training recipe and distillation strategy to another MLLM framework MobileVLM [1].

B.4. Ablation study on teacher models with different sizes

Table A8 shows more detailed results of the ablation study that investigates the impact of different teacher scales.

C. Implementation Details

Table A9 presents the training hyperparameters for LLaVA-KD across different learning phases. During the Distilled

Response Tokens	Prompt Tokens	Visual Tokens	Image Question Answering					Benchmarks					Avg_{10}
			VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
✓	✗	✗	73.8	57.8	25.6	62.8	47.1	59.7	55.9	49.3	85.5	31.6	54.9
✓	✓	✗	74.1	58.2	24.4	60.6	48.6	59.9	56.3	50.6	84.8	32.3	55.0
✓	✗	✓	74.5	58.3	26.7	62.6	48.5	57.3	57.1	48.6	85.6	31.8	55.1
✓	✓	✓	74.2	58.3	24.6	60.4	46.9	60.0	55.6	49.1	84.8	32.2	54.6

(a) Distillation targets during the DPT stage.

Response Tokens	Prompt Tokens	Visual Tokens	Image Question Answering					Benchmarks					Avg_{10}
			VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
✓	✗	✗	76.8	59.6	36.4	59.1	50.2	64.0	57.6	52.7	85.8	30.1	57.2
✓	✓	✗	77.0	59.5	27.5	60.1	51.5	62.7	59.5	55.8	85.7	30.0	56.9
✓	✗	✓	76.9	59.7	38.3	59.9	49.4	64.1	57.4	54.8	86.3	30.7	57.7
✓	✓	✓	76.4	59.0	30.8	61.4	49.9	63.5	59.2	55.1	86.0	29.9	57.1

(b) Distillation targets during the DFT stage.

Table A6. Ablation studies on distillation targets during DPT and DFT stages.

LLM of the Teacher	LLM of the Student	Training Recipe	Image Question Answering					Benchmarks					Avg_{10}
			VQAv2	GQA	VisWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
MLLaMA 2.7B	/	PT-SFT	73.9	58.0	32.7	60.6	46.9	65.0	60.3	37.9	84.4	31.9	55.2
/	MLLaMA 1.7B	PT-SFT	70.1	55.2	24.7	58.3	40.5	58.1	50.3	16.8	84	29.9	48.8
MLLaMA 2.7B	MLLaMA 1.7B	DPT-SFT	71.1	56.5	30.7	57.2	41.2	58.3	52.3	22.9	84.8	30.1	50.5
MLLaMA 2.7B	MLLaMA 1.7B	DPT-SFT-DFT	72.7	57.2	40.0	58.6	41.3	62.4	54.1	32.0	85.0	30.3	53.4

Table A7. Detailed results of verification on MobileVLM.

LLM of the Teacher	LLM of the Student	Image Question Answering					Benchmarks					Avg_{10}
		VQAv2	GQA	VisWiz	SciQA	TextVQA	MME	MMB	MMB-CN	POPE	MMMU	
Qwen1.5-4B	/	79.9	63.4	46.3	72.9	59.0	69.3	67.9	67.1	85.2	38.9	65.0
Qwen1.5-7B	/	80.5	63.3	48.6	70.1	58.7	70.9	70.5	68.7	86.8	39.2	65.7
/	Qwen1.5-0.5B	73.9	57.4	24.9	60.9	47.4	59.8	55.0	52.4	83.7	31.6	54.7
Qwen1.5-4B	Qwen1.5-0.5B	77.0	59.6	35.9	60.6	49.9	64.5	60.1	55.5	85.9	30.2	57.9
Qwen1.5-7B	Qwen1.5-0.5B	76.9	59.1	35.9	59.5	49.3	63.1	58.0	54.4	86.6	31.4	57.4
Qwen2.5-3B	/	80.4	63.2	38.7	76.0	61.5	73.9	71.8	69.5	86.4	40.3	66.2
Qwen2.5-7B	/	81.8	64.3	46.1	77.3	64.6	78.5	74.7	73.0	86.6	46.4	69.3
/	Qwen2.5-0.5B	74.8	58.3	28.9	59.1	49.2	61.5	58.9	54.2	86.1	33.6	56.5
Qwen2.5-3B	Qwen2.5-0.5B	77.7	59.8	41.5	60.6	52.0	64.7	61.3	57.0	86.4	28.3	58.9
Qwen2.5-7B	Qwen2.5-0.5B	77.6	59.6	39.9	61.0	51.2	62.6	57.9	56.3	86.2	30.8	58.3

Table A8. Ablation study on teacher models with different sizes.

Hyperparameter	DPT	SFT	DFT
Visual Encoder	✗	✗	✗
Projector	✓	✓	✓
LLM	✗	✓	✓
Image Resolution	384×384		
Learning Rate	1.00e-03	2.00e-05	2.00e-05
Optimizer	AdamW		
Scheduler	Cosine decacy		
Warm up ratio	0.03		
Global Batch Size	256	128	128
Epoch	1		
DeepSpeed stage	Zero 2	Zero 2	Zero 3

Table A9. Hyperparameters of LLaVA-KD.

Pre-Training (DPT) stage, we exclusively optimize the Projector network to align the visual and textual modalities. During the followed Supervised Fine-Tuning (SFT) and Distilled Fine-Tuning (DFT) stages, we jointly train both the Projector and LLM to enhance the model's multimodal understanding capabilities.

References

[1] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 3

[2] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. 1

[3] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 2

[4] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 1

[5] Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, Lihao Zheng, Zhenbiao Gai, Mingyang Wang, and Jiajun Ding. Imp: Highly capable large multimodal models for mobile devices. *arXiv preprint arXiv:2405.12107*, 2024. 2

[6] Fangxun Shu, Yue Liao, Le Zhuo, Chenning Xu, Guang-

hao Zhang, Haonan Shi, Long Chen, Tao Zhong, Wanggui He, Siming Fu, et al. Llava-mod: Making llava tiny via moe knowledge distillation. *arXiv preprint arXiv:2408.15881*, 2024. 1, 2

[7] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 2, 3