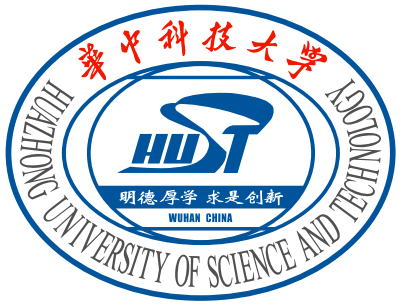
ICCV OCT 19-23, 2025 HONOLULU HAWAII

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

Yuxuan Cai*, Jiangning Zhang*, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, Zhucun Xue, Yong Liu, Xiang Bai†

Huazhong University of Science and Technology, Zhejiang University

Tencent Youtu Lab, Huazhong Agricultural University



Background

Background

Multimodal Large Language Models (MLLMs) exhibit remarkable performance but face significant deployment challenges:

- **Massive Parameter Sizes:** Enormous computational requirements
- **High Inference Costs:** Resource-intensive operations
- **Resource-Constrained Environments:** Limited computational resources in practical deployment settings

Core Problem

How to effectively compress and streamline MLLMs to reduce computational overhead while maintaining competitive performance?

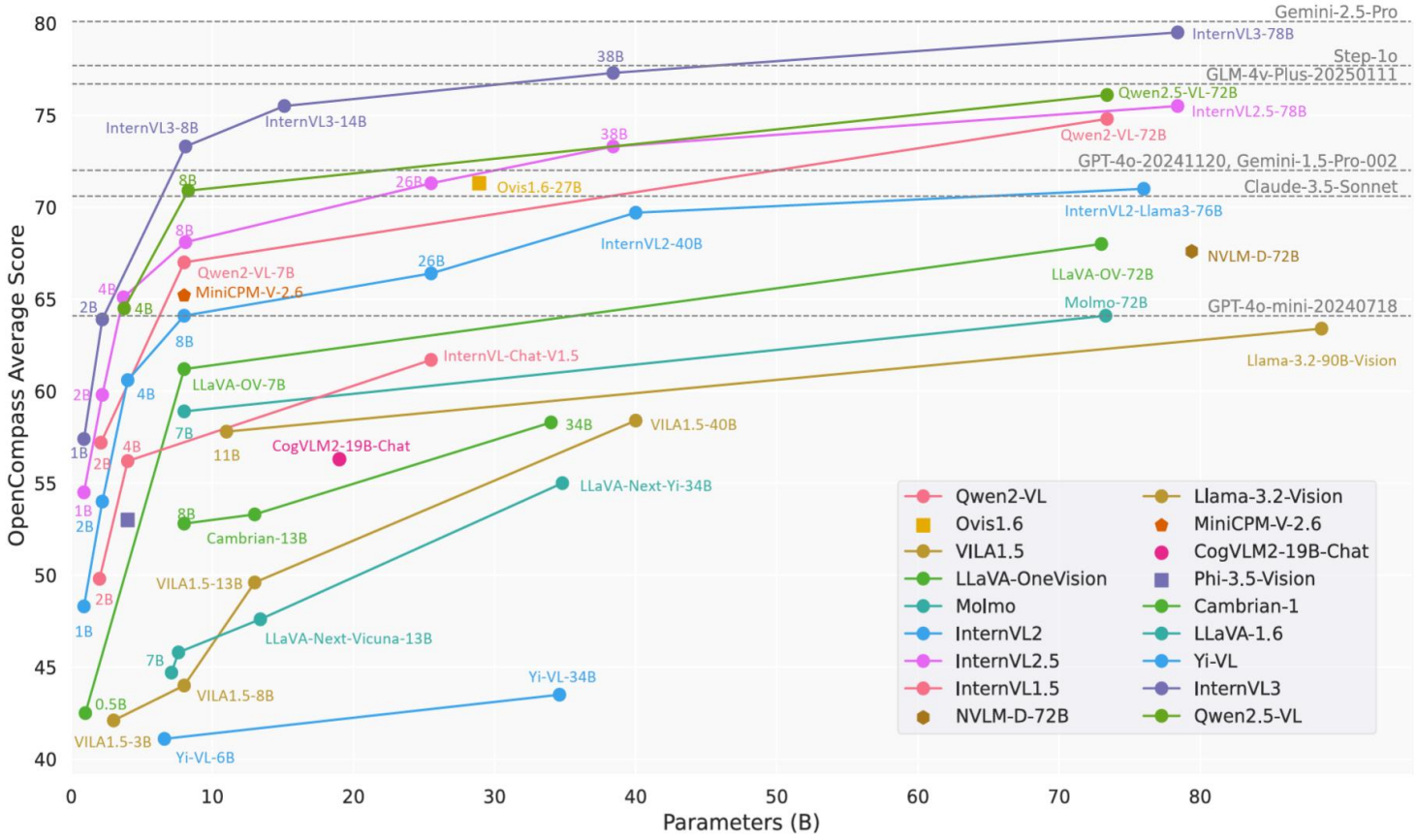


Image source: *InternVL3.0 Technical Report*

Motivation

Model Capacity Limitation

- Small-scale MLLMs struggle to learn complex cross-modal knowledge
- Conventional PT and SFT paradigm insufficient
- Fundamental capacity gap requires sophisticated distillation

Distillation Timing Issue

- Knowledge distillation applied only during SFT stage
- Limited improvements from isolated distillation
- Missing systematic training paradigm throughout learning

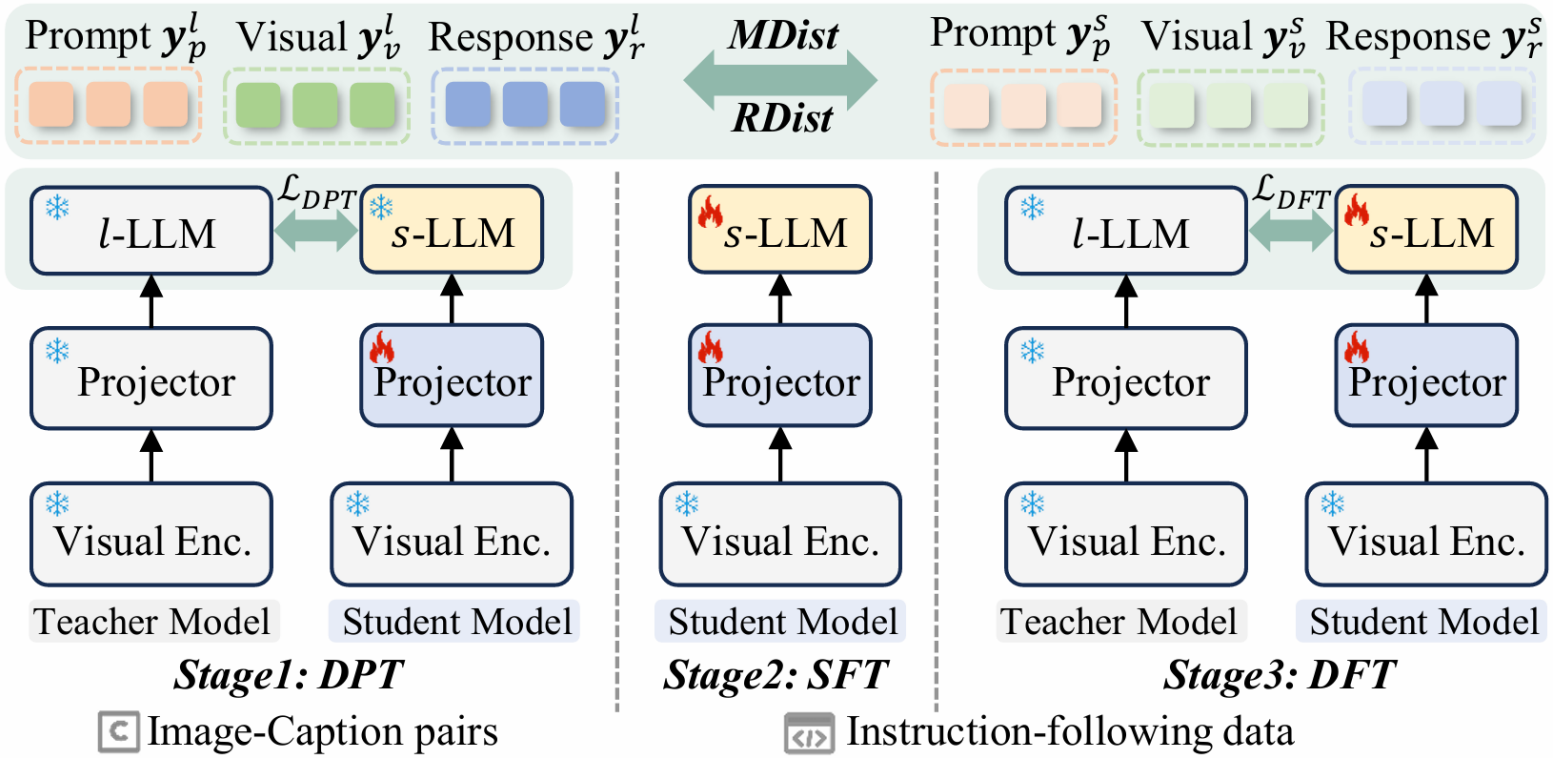
Single Distillation Objective

- Focus on textual responses, overlooking visual modality
- Visual relationships between objects not captured
- Limited comprehensive transfer of multimodal knowledge

◎ **Goal:** Develop a training paradigm tailored for small-scale MLLMs that enhances multimodal understanding

Method

Three-Stage Distillation Training Pipeline



Distilled Pre-Training (DPT)

Purpose: Enhances alignment between visual and linguistic representations

Benefit: Lays foundation for cross-modal understanding

Supervised Fine-Tuning (SFT)

Purpose: Equips student model with multimodal reasoning capabilities

Benefit: Enables learning from human-annotated data

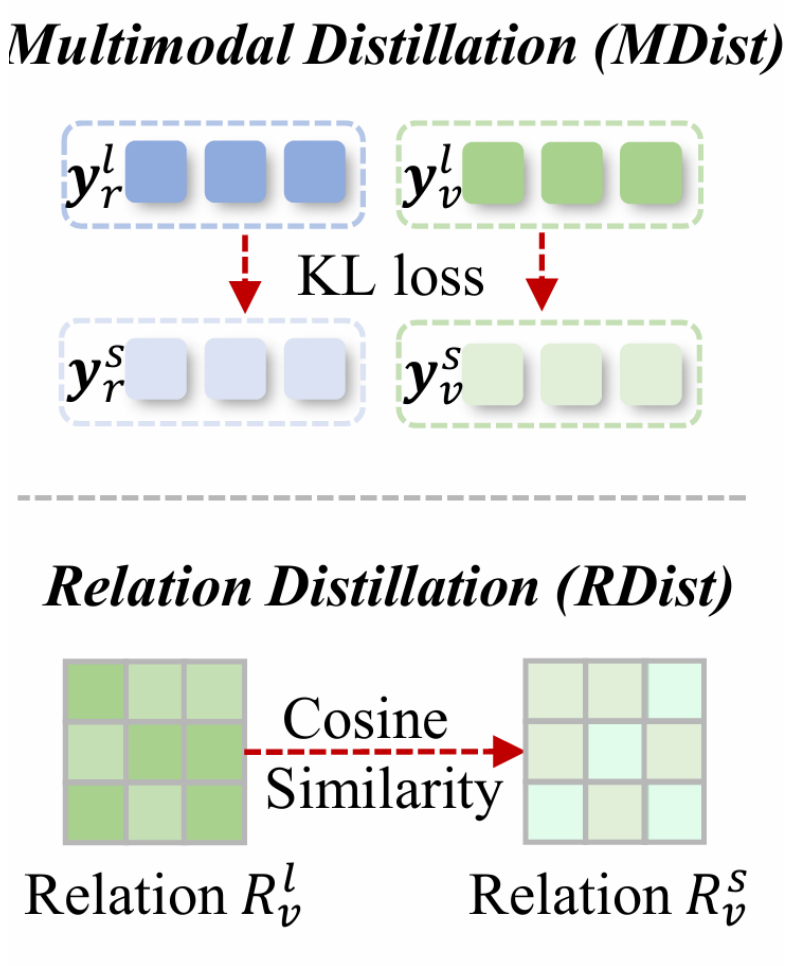
Distilled Fine-Tuning (DFT)

Purpose: Refines student's multimodal comprehension

Benefit: Enhances deep reasoning ability

Method

Training Strategies



Multimodal Distillation (MDist)

Mechanism

Distills the output distributions of visual tokens from the teacher model to the student model

Objective

Ensures that the student model captures fine-grained visual information, such as object features and local details, learned by the teacher

Relational Distillation (RDist)

Mechanism

Constructs self-correlation matrices from visual tokens of the teacher model and transfers this relational knowledge to the student

Objective

Transfers the teacher's ability to model complex visual relationships, including spatial arrangements and interactions between objects within a scene

Experiments

Benchmarked results with SoTA MLLMs

Method	LLM	#Samples	Image Question Answering					Benchmarks					Avg_7	Avg_{10}
			VQAv2	GQA	VisWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU		
LLaVA-1.5	Vicuna-7B	1.2 M	78.5	62.0	50.0	66.8	58.2	75.5	64.3	58.3	85.9	34.4	62.2	63.4
InstructBLIP	Vicuna-7B	130 M	-	49.2	-	60.5	50.1	-	36.0	-	-	-	-	-
Qwen-VL	Qwen-7B	1500 M	78.8	59.3	35.2	67.1	63.8	-	38.2	7.4	-	-	-	-
Qwen-VL-Chat	Qwen-7B	1500 M	78.2	57.5	38.9	68.2	61.5	74.4	60.6	56.7	-	35.9	59.7	-
mPLUG-Owl2	LLaMA2-7B	400 M	79.4	56.1	54.5	68.7	54.3	72.5	66.5	-	85.8	32.7	62.1	-
TinyLLaVA†	Qwen1.5-4B	1.2 M	79.9	63.4	46.3	72.9	59.0	69.3	67.9	67.1	85.2	38.9	63.7	65.0
TinyLLaVA†	Qwen2.5-3B	1.2 M	80.4	63.2	38.7	76.0	61.5	73.9	71.8	69.5	86.4	40.3	64.9	66.2
TinyLLaVA	Phi2-2.7B	1.2 M	79.9	62.0	-	69.1	59.1	73.2	66.9	-	86.4	38.4	-	-
Bunny	Phi2-2.7B	2.6 M	79.8	62.5	43.8	70.9	56.7	74.4	68.6	37.2	-	38.2	59.2	-
Imp-3B	Phi2-2.7B	1.5 M	-	63.5	54.1	72.8	59.8	-	72.9	46.7	-	-	-	-
MobileVLM	MLLaMA-2.7B	1.2 M	-	59.0	-	61.0	47.5	64.4	59.6	-	84.9	-	-	-
MoE-LLaVA	Phi2-2.7B	2.2 M	79.9	62.6	-	70.3	57.0	-	68.0	-	85.7	-	-	-
MiniCPM-V	MiniCPM-2.4B	570 M	-	51.5	50.5	74.4	56.6	68.9	64.0	62.7	79.5	-	61.2	-
LLaVADI	MLLaMA-2.7B	1.2 M	-	61.4	-	64.1	50.7	68.8	62.5	-	86.7	-	-	-
Imp-2B	Qwen1.5-1.8B	1.5 M	79.2	61.9	39.6	66.1	54.5	65.2	63.8	61.3	86.7	-	58.9	-
Bunny-2B	Qwen1.5-1.8B	2.6 M	76.6	59.6	34.2	64.6	53.2	65.0	59.1	58.5	85.8	-	56.3	-
Mini-Gemini-2B	Gemma-2B	2.7 M	-	60.7	41.5	63.1	56.2	67.0	59.8	51.3	85.6	31.7	57.1	-
MoE-LLaVA-2B	Qwen-1.5-1.8B	2.2 M	76.2	61.5	32.6	63.1	48.0	64.6	59.7	57.3	**87.0**	-	55.3	-
TinyLLaVA†	Qwen2.5-1.5B	1.2 M	78.8	62.0	43.2	**72.0**	57.4	**72.5**	68.6	63.0	85.5	37.0	62.7	64.0
TinyLLaVA†	Qwen1.5-1.8B	1.2 M	73.1	55.5	34.9	65.3	47.7	61.2	57.1	55.5	83.4	34.1	53.9	56.8
LLaVA-MOD	Qwen1.5-1.8B	5 M	-	58.7	39.2	68.0	58.5	66.7	66.3	61.9	**87.0**	-	59.9	-
LLaVA-KD	Qwen1.5-1.8B	1.2 M	79.0	62.3	44.7	64.7	53.4	69.1	64.0	63.7	86.3	33.6	60.3	62.1
LLaVA-KD	Qwen2.5-1.5B	1.2 M	**80.3**	**62.5**	**46.0**	71.6	**59.7**	**70.0**	**71.0**	**66.6**	86.7	**35.8**	**63.9**	**65.0**
SPHINX-Tiny	TinyLlama-1.1B	15 M	74.7	58.0	49.2	21.5	57.8	63.1	52.3	56.6	82.2	-	51.2	-
TinyLLaVA†	Qwen1.5-0.5B	1.2 M	73.9	57.4	24.9	60.9	47.4	59.8	55.0	52.4	83.7	31.6	51.1	54.7
TinyLLaVA†	Qwen2.5-0.5B	1.2 M	74.8	58.3	28.9	59.1	49.2	61.5	58.9	54.2	86.1	33.6	52.9	56.5
LLaVADI	MLLaMA-1.4B	1.2 M	-	55.4	-	56.0	45.3	58.9	55.0	-	84.7	-	-	-
LLaVA-MOD	Qwen1.5-0.5B	5 M	-	56.2	31.6	62.8	53.9	65.3	58.8	50.4	-	-	54.1	-
LLaVA-KD	Qwen1.5-0.5B	1.2 M	77.0	59.6	35.9	60.6	49.9	64.5	60.1	55.5	85.9	30.2	55.2	57.9
LLaVA-KD	Qwen2.5-0.5B	1.2 M	**77.7**	**59.8**	41.5	60.6	52.0	64.7	**61.3**	**57.0**	**86.4**	28.3	**56.7**	**58.9**

LLaVA-KD consistently outperforms existing small-scale MLLMs across multiple standard multimodal benchmarks.

Experiments

Visualization Comparison with Baseline

Experiments

Ablation studies of different training stages

Training Scheme	Image Question Answering					Benchmarks					Avg_{10}
	VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
PT-SFT	73.9	57.4	24.9	60.9	47.4	59.8	55.0	52.4	83.7	**31.6**	54.7
DPT-SFT	74.6	57.8	28.8	**61.2**	49.1	59.9	56.9	51.6	84.3	31.4	55.6
PT-DFT	75.1	57.0	29.5	60.9	49.2	59.6	57.3	55.0	85.5	29.6	55.8
DPT-DFT	75.5	58.0	27.5	59.7	49.3	60.6	57.7	54.7	85.4	30.3	55.9
PT-SFT-DFT	76.6	59.4	32.6	60.4	48.4	60.9	57.8	54.0	84.9	31.3	56.6
DPT-SFT-DFT	**77.0**	**59.6**	**35.9**	60.6	**49.9**	**64.5**	**60.1**	**55.5**	**85.9**	30.2	**57.9**
DPT-DFT-DFT	77.5	60.3	37.6	61.1	49.3	63.2	59.4	54.9	86.0	31.0	58.0

↔ DPT improves visual-text alignment

DPT enhances visual–text alignment and multimodal understanding between representations

DFT refines understanding

DFT enhances knowledge transfer and refines the student's multimodal comprehension

SFT enables effective learning

SFT is essential for the student model to learn from supervised human-annotated data

Three-stage pipeline performs best

DPT-SFT-DFT achieves optimal balance with highest average scores across benchmarks

Ablation Study on MDist and RDist

DPT		SFT	DFT		Avg_{10}
MDist	RDist		MDist	RDist	
✗	✓		✗	✗	55.5
✓	✗	✓	✗	✗	55.1
✓	✓		✗	✗	**55.6**
✓	✓		✗	✓	57.0
✓	✓	✓	✓	✗	57.7
✓	✓		✓	✓	**57.9**

- Combined MDist and RDist across both stages yields optimal performance
- MDist focuses on fine-grained visual information (object features)
- RDist captures complex visual relationships (spatial arrangements)

Experiments

Comparison with Distillation Strategies in LLMs

Distillation strategy	Image Question Answering					Benchmarks					Avg_{10}
	VQAv2	GQA	VizWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
FKL	74.3	56.1	31.7	59.4	49.0	58.9	57.4	54.0	84.4	29.8	55.5
RKL [11]	74.3	56.6	26.7	60.8	49.1	57.8	56.8	53.7	84.7	**30.0**	55.0
JSD [38]	73.8	54.9	**32.3**	60.3	48.7	57.6	**57.8**	54.3	85.1	29.8	55.5
Ours	**75.1**	**57.0**	29.5	**60.9**	**49.2**	**59.6**	57.3	**55.0**	**85.5**	29.6	**55.8**

Results indicate that specialized multimodal distillation strategies outperform generic LLM distillation methods for MLLMs.

Comparison with distillation pipeline in LLaVA-MOD

Distillation Strategy	Image Question Answering					Benchmarks					Avg_{10}
	VQAv2	GQA	VisWiz	SciQA	TextVQA	MME	MMB	MMB^{CN}	POPE	MMMU	
MD	76.3	58.5	31.6	58.4	**51.7**	60.6	59.6	**55.8**	86.2	30.2	56.9
MD+PD	74.4	57.1	22.7	58.4	47.7	58.4	58.8	54.5	**86.6**	**32.1**	55.1
Ours	**77.0**	**59.6**	**35.9**	**60.6**	49.9	**64.5**	**60.1**	55.5	85.9	30.2	**57.9**

Compared with recent distillation-based methods such as LLaVA-MoD, LLaVA-KD achieves superior performance using significantly less training data (1.2M vs. 5M), surpassing LLaVA-MoD by approximately +1.1%.

Conclusion

Three-Stage Training Pipeline

Systematically integrates distillation across different training phases to ensure comprehensive knowledge transfer from large teacher MLLMs to small student MLLMs.

- Distilled Pre-Training (DPT): Enhances visual–language representation alignment
- Supervised Fine-Tuning (SFT): Equips multimodal reasoning and instruction-following capabilities

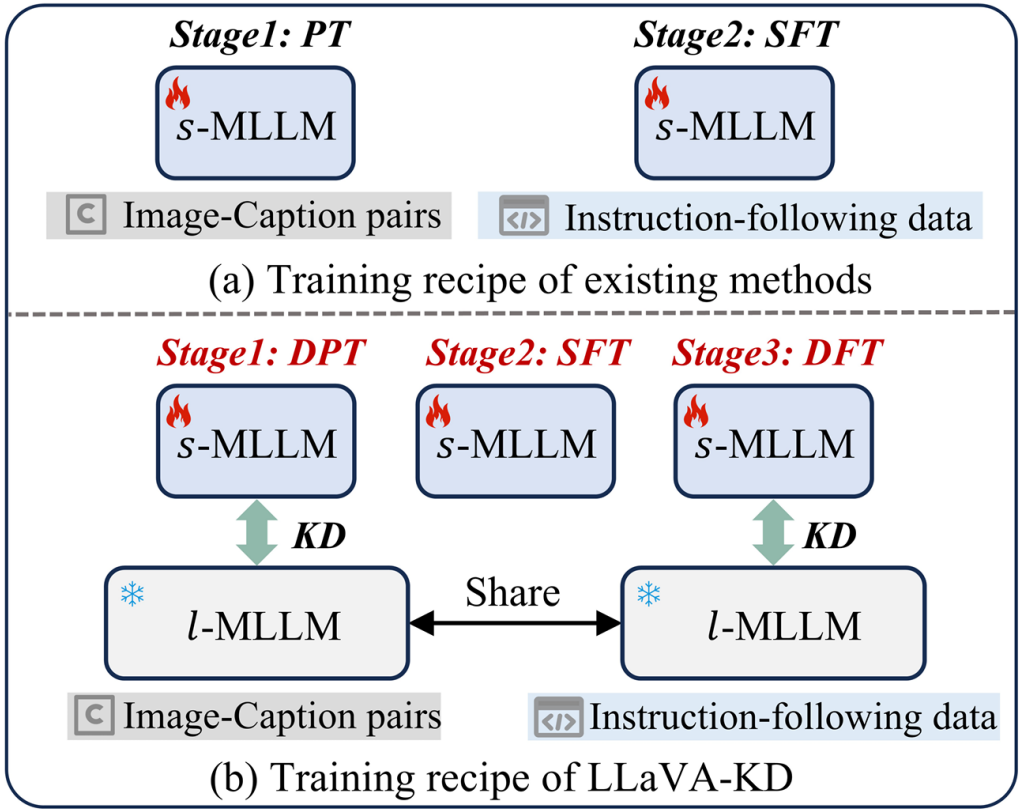
Dual Modal Distillation Design

Introduces two distinct distillation mechanisms to capture both fine-grained visual information and complex visual relationships.

- Multimodal Distillation (MDist): Captures fine-grained visual information
- Relational Distillation (RDist): Transfers complex visual relationships and interactions

Strong Empirical Results

Consistent improvements across multiple benchmarks, demonstrating superior performance and efficiency compared to existing methods.



THANK YOU

https://github.com/Fantasyele/LLaVA-KD