LLaVA-KD ICCV 2025

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

하동윤

2025.11.14

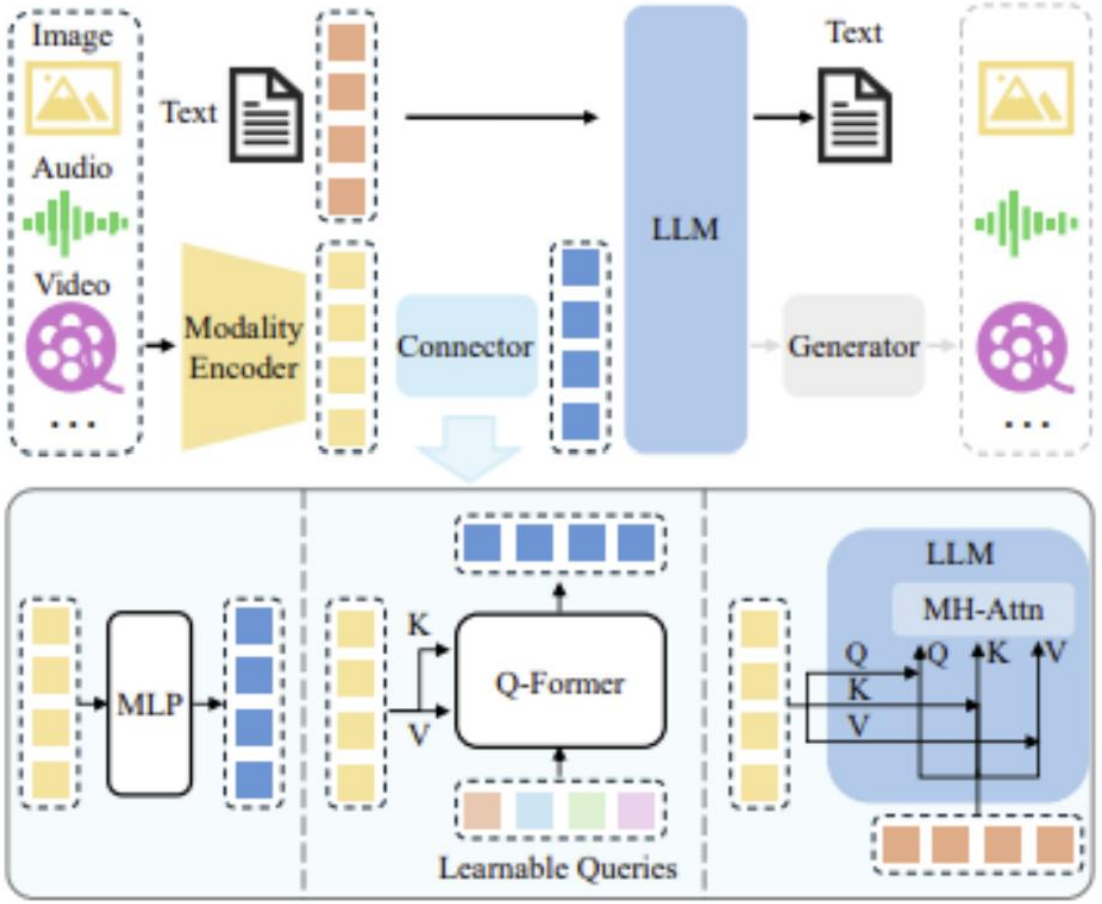
GSPS
Graduate Student Paper Seminar

What is MLLM?

- 기존의 LLM은 이미지를 못받고 LVM은 추론 과정에서 지연이 생김
  → LLM + LVM = MLLM

(multimodal: 단일 입력이 아닌 2개 이상의 입력이 들어갈 때 표현하는 언어)

- 현대 MLLM의 대표적인 베이스라인 모델로는 LLaVA가 있습니다



What is LLaVA?

이미지 + 언어 통합 모델(MLLM)을 Instruction Tuning 방식으로 학습시켜 대화, 묘사, 추론이 가능하도록

Visual Encoder: CLIP Vit(이미지 -> 시각 feature)

Projection Layer로 LLM의 Embedding 공간에 정렬(LLaVA를 가장 먼저 제안한 논문에선 Single Layer로 단순히 붙이는 방식에 그쳤지만 후속 논문인 LLaVA 1.5에선 MLP(multi layer perceptron)을 활용해 개선

이후에 Vicuna라는 LLM에 이미지 정보를 텍스트 토큰과 같게 활용

What is LLaVA?

Pre-training : Vision Encoder & LLM을 Freeze하여 projectio만 학습 -> 시각-언어 공간 정렬

Fine-tuning : Encoder Freeze / Projection + LLM 업데이트

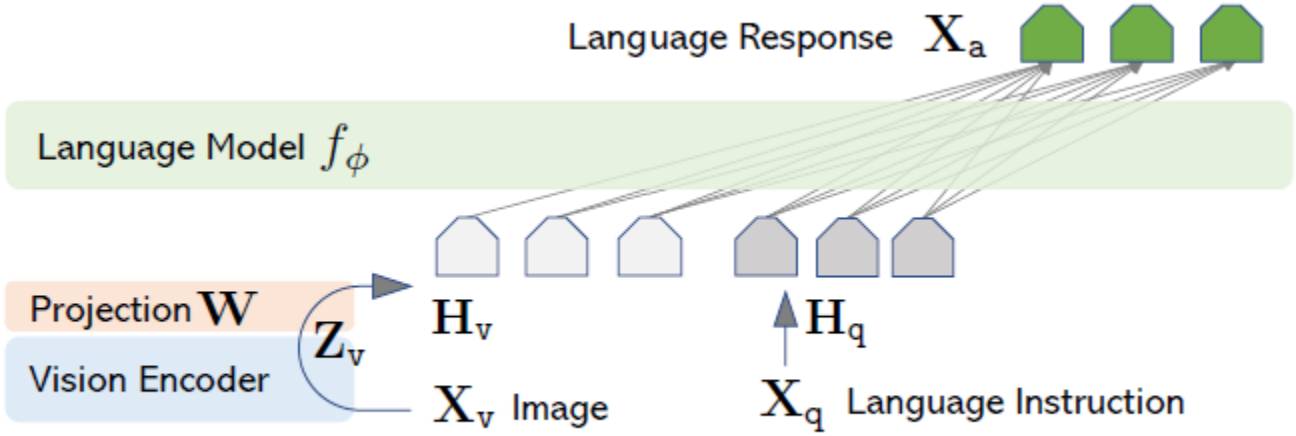


Figure 1: LLaVA network architecture.

What is KD(Knowledge Distillation)?

거대한 teacher model에서 lightweight student 모델로 지식을 이전 → fewer parameters, less computation, and faster speed

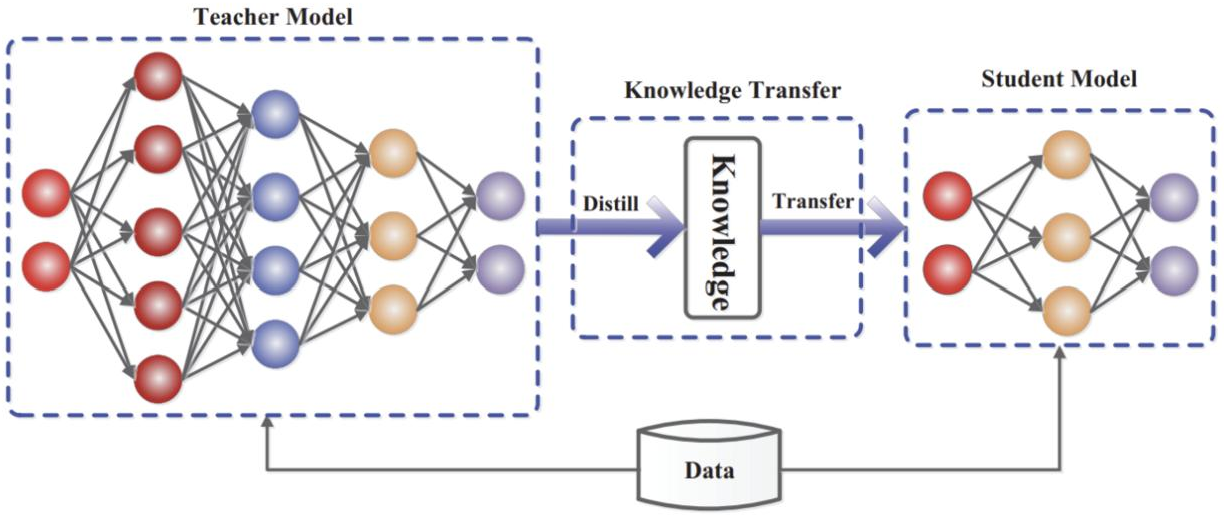


Image Credit: 'Knowledge Distillation: A Survey' 논문

What is KD(Knowledge Distillation)?

응답 (Response) 기반의 증류 (‘출력값’이 지식)
교사 모델의 최종적인 출력값/확률을 학생 모델을 훈련하는 목표 대상으로 훈련 -> 이미지 분류, 객체 탐지, 자세 예측 같은 다양한 작업을 대상으로 사용합니다.

피처 (Feature) 기반의 증류 (‘중간 계층’이 지식)
결과값, 즉 ‘최종 예측’을 복사하는게 아니라, 학생 모델이 교사 모델의 중간 계층, 피처 맵 등을 통해서도 학습 -> 어텐션 맵, 확률 분포, 계층 연결 같은 다양한 방법들을 통해서 교사 모델과 학생 모델의 피처 (Feature)를 일치시키는데 도움을 줄 수 있고, 특히 이미지 인식, 객체 탐지 같은 작업의 성능을 향상시켜 준다고 합니다.

What is KD(Knowledge Distillation)?

관계 (Relationships) 기반의 증류 (‘관계’가 지식)
학생 모델이, 교사 모델 내부의 여러 부분들 사이의 관계를 모방하는 법을 학습 - 계층 간, 또는 다른 데이터 샘플 간의 관계

⇒ 기존 Distillation은 대부분 텍스트 출력만 맞추는 방식

→ 이미지 정보(Visual Token) 관계 이해 능력은 제대로 전이되지 않음

How does Knowledge Distillation work?

: with Hinton's KD

- Soft Label: 일반적으로, 이미지 클래스 분류와 같은 task는 신경망의 마지막 softmax layer를 통해 각 클래스의 확률값을 뱉어내게 됩니다.(아래는 i번째 클래스에 대한 확률값 q_i, z_i는 모델이 출력한 점수 또는 logit)

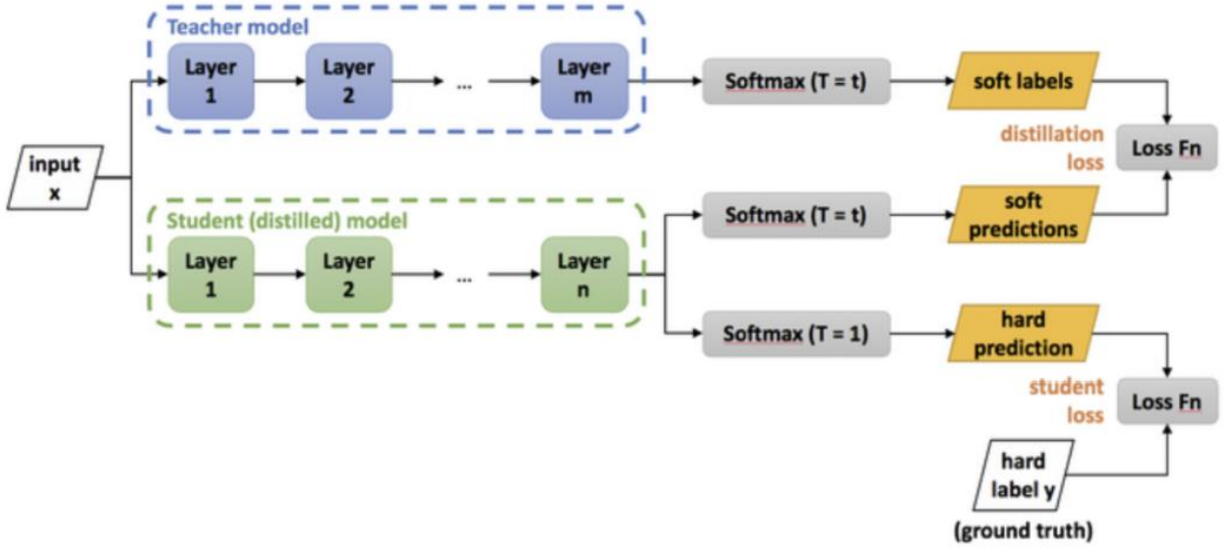
$$q_i = \frac{exp(z_i)}{\sum_j exp(z_j)}$$

- 출력값의 분포를 좀 더 soft하게 만들면, 이 값들의 모델이 가진 지식이라 볼 수 있다. 아래의 temperature에 해당한 T가 높아지면 확률 분포가 납작해지고 클래스간 관계 정보를 더 명확하게 나타냄

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

How does Knowledge Distillation work?

: with Hinton's KD



$$L = \sum_{(x,y)\in\mathbb{D}} L_{KD}(S(x,\theta_S,\tau), T(x,\theta_T,\tau)) + \lambda L_{CE}(\hat{y}_S, y)$$

- L: 손실함수
- S: Student Model
- T: Teacher Model
- (x, y): 하나의 이미지와 그 레이블
- θ: 모델의 학습 파라미터
- τ: temperature
- Lce: cross entropy loss

Composition of LLaVA-KD

: Frozen Visual Encoder & Visual Projector

- Frozen Visual Encoder
  - Input image X_v → sequenced to 2D patches Pv

$$P_v \in \mathbb{R}^{N_p \times S_p^2 \times 3}$$

    - S_p: patch size
    - N_p: patch number
  - → transformer layer가 visual feature로 project

$$Z_v \in \mathbb{R}^{N_p \times C},$$

    - C: dimension
  - Visual encoder는 pre-trained SigLIP

- Visual Projector
  - 2개의 MLP layer + GELU → Z_v(visual feature)를 H_v(text embedding space)에 project

$$H_v \in \mathbb{R}^{N_p \times D},$$

    - D: embedding dimension

Composition of LLaVA-KD
: Large Language Model

- Large Language Model
  - multimodal information 합침(H_v: visual embedding, H_t: text embedding)

$$H = [H_v, H_t]$$

  - generate sequential output

$$\mathbf{y} = [\mathbf{y}_p, \mathbf{y}_v, \mathbf{y}_r] = \{y_t\}_{t=1}^{T}$$

    - y_p: prompt
    - y_v: visual
    - y_r: response
    - T: prediction token의 길이

Teacher Model의 Scheme

Pre-training

- Visual Encoder과 LLM은 frozen
- Projector만 최적화(visual과 textual 정렬과정)

$$\mathcal{L}_{reg} = -\sum_{m=1}^{M} \log \phi_l \left(y_m \mid \mathbf{y}_{<m} \right),$$

- image-caption 쌍으로 위의 목적을 달성
  - M: response token의 예상길이
  - y_m: predicted reponse token
  - $y_{<m}$: previous prediction

  $\phi_l \left(y_m \mid \mathbf{y}_{<m} \right)$

  = 이전의 예측을 바탕으로 현재의 response token의 재정렬

Teacher Model의 Scheme

Supervised Fine-Tuning

- visual encoder frozen
- projector와 LLM의 최적화
- 높은 퀄리티의 instruction dataset으로 objective function

$\mathcal{L}_{SFT}$ 를 따름(이전의 공식처럼 autoregressive하게)

MLLM-Oriented KD Strategy(for s-MLLM)

Multimodal Distillation(MDist) - 어차피 multimodal도 LLM으로 이해후 추론하기에 기존의 distillation 방식 사용: standard KLD(Kullback-Leibler Divergence)

$$\mathcal{L}_{res} = \sum_{m=1}^{M} \mathrm{KLD}(\phi_l(y_m \mid \mathbf{y}_{<m}), \phi_s(y_m \mid \mathbf{y}_{<m})),$$

$$= \sum_{m=1}^{M} \sum_{j=1}^{V} \phi_l\left(Y_j \mid \mathbf{y}_{<m}\right) \log \left(\frac{\phi_l\left(Y_j \mid \mathbf{y}_{<m}\right)}{\phi_s\left(Y_j \mid \mathbf{y}_{<m}\right)} \right),$$

- M: response token의 길이
- V: vocabulary space
- ϕ_l: l-MLLM의 parameter
- ϕ_s: s-MLLM의 parameter
- Y_j: vocabulary
- y_m: response token
- $\phi_l(Y_j|y < m)$: response token에서 단어의 확률 by l-MLLM
- $\phi_s(Y_j|y < m)$: response token에서 단어의 확률 by s-MLLM

What is KLD?

- Kullback-Leibler Divergence: 두 확률분포의 차이를 계산하는 데에 사용하는 함수로, 어떤 이상적인 분포에 대해, 그 분포를 근사하는 다른 분포를 사용해 샘플링을 한다면 발생할 수 있는 정보 엔트로피 차이를 계산한다.

$$KL(p \parallel q) = \begin{cases} \sum_i p_i \log \frac{p_i}{q_i} \text{ 또는 } -\sum_i p_i \log \frac{q_i}{p_i} & \text{(이산형)} \\ \int p(x) \log \frac{p(x)}{q(x)} dx \text{ 또는 } -\int p(x) \log \frac{q(x)}{p(x)} dx & \text{(연속형)} \end{cases}$$

- 예시: 동전이 있다고 할 때 실제 동전(P)은 앞면이 0.7, 뒷면이 0.3이라 할 때

KL(P||Q) = 0.7log(0.7/0.5) + 0.3log((0.3/0.5) -> 앞면 과소추정하여 비용증가 / 뒷면 과대 추정하여 비용 증가

KLD는 참 분포 P를 모델 Q로 설명하려고 할 때 얼마나 비효율적인지 측정하는 값

MDist continued

$$\mathcal{L}_{vis} = \sum_{k=1}^{K} \sum_{j=1}^{V} \phi_l \left(Y_j \mid \mathbf{y}_{<k} \right) \log \left(\frac{\phi_l \left(Y_j \mid \mathbf{y}_{<k} \right)}{\phi_s \left(Y_j \mid \mathbf{y}_{<k} \right)} \right),$$

vision도 신경써야 하기에

K = visual token

Relation Distillation(RDist)

relational reasoning 능력을 위해 self-correlation matrix(LLM에서 생성된 visual token 활용)

$$\begin{cases} R_v^s = \mathbf{y}_v^s \otimes \mathbf{y}_v^s \in \mathbb{R}^{N_p \times N_p}, \\ R_v^t = \mathbf{y}_v^t \otimes \mathbf{y}_v^t \in \mathbb{R}^{N_p \times N_p}, \end{cases}$$

- $\otimes$: matrix multiplication
- y_v^s: student의 visual token
- y_v^t: teacher의 visual token
- N_p: number of visual token

$$\mathcal{L}_{rel} = 1 - \mathrm{Cos}(R_v^s, R_v^t) = 1 - \frac{R_v^s \cdot R_v^t}{\|R_v^s\| \|R_v^t\|},$$

Distillation Scheme

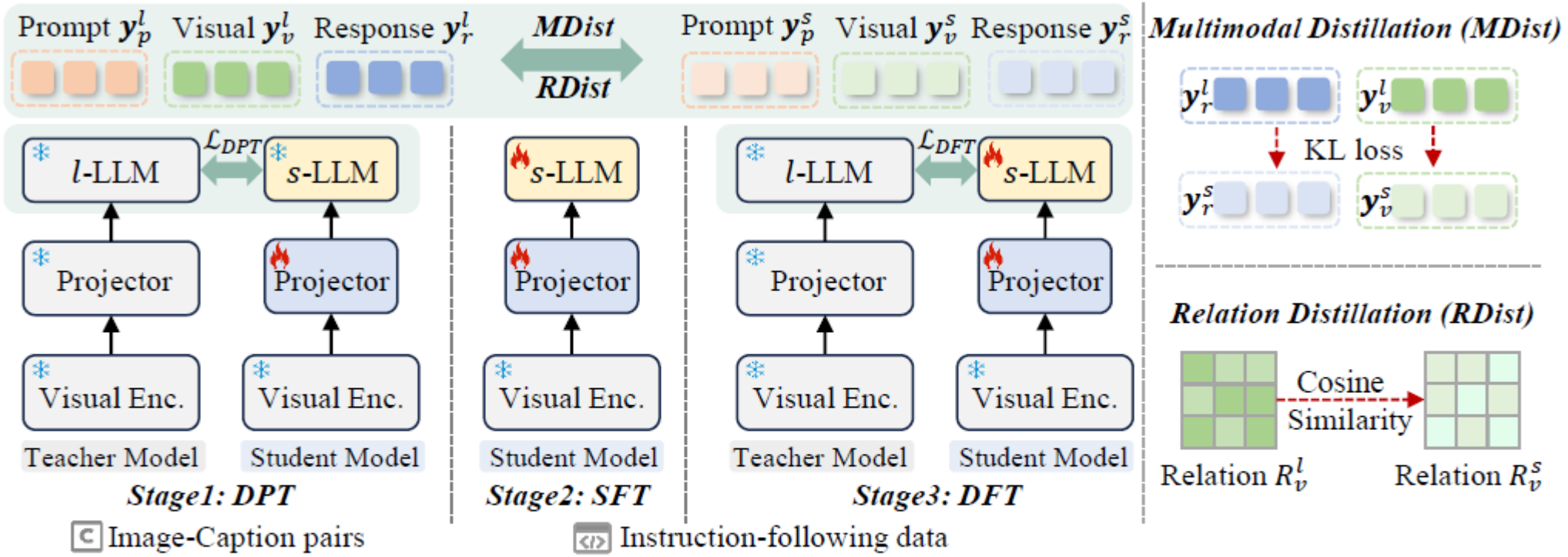


Figure 2. **Overview of our LLaVA-KD** that contains three stages for effect training: *1)* Distilled Pre-Training (DPT) to align visual and text information as l-MLLM. *2)* Supervised Fine-Tuning (SFT) to enable s-MLLM with multimodal understanding capacity. *3)* Distilled Fine-Tuning (DFT) to refine s-MLLM's capacity by transferring l-MLLM's knowledge. During the training phase, we employ Multimodal Distillation (MDist) and Relation Distillation (RDist) in both DPT and DFT stages.

Distillation Scheme

Distilled Pre-Training: $\mathcal{L}_{DPT} = \mathcal{L}_{reg} + \alpha\mathcal{L}_{res} + \beta\mathcal{L}_{vis} + \gamma\mathcal{L}_{rel},$

Supervised Fine-tuning

Distilled Fine-Tuning: $\mathcal{L}_{DFT} = \mathcal{L}_{reg} + \alpha'\mathcal{L}_{res} + \beta'\mathcal{L}_{vis} + \gamma'\mathcal{L}_{rel},$

단계	이름	목적	Distillation 적용 여부
1	**DPT (Distilled Pre-Training)**	시각-텍스트 정렬 강화	**MDist + RDist 사용**
2	**SFT (Supervised Fine-Tuning)**	기본적인 instruction-following 능력 부여	Distillation **사용 안 함**
3	**DFT (Distilled Fine-Tuning)**	Teacher와 최대한 비슷하게 완성도 높이기 (multimodal reasoning 강화)	**MDist + RDist 다시 적용**

Experiments

- Visual Encoder: SigLIP-B/14@384 사용(teacher와 Student 모두 동일하게 -> distillation 과정에서 visual varianc를 줄이기 위해)
- LLM Backbone(Vocab, tokenizer 체계가 일치해야 함))
  - Student: Qwen 1.5/2.5(0.5B, 1.8B 등)
  - Teacher: 4B, 7B 등의 Qwen 계열
- Training Datasets
  - DPT: LLaVA 1.5 – 558k Image-Caption 데이터
  - SFT, DFT: LLaVA-mix 665k instruction 데이터
- Loss weight(visual 관계 정보의 중요성은 인정하지만 너무 크게 주진 않음)
  - MDist(res/vis) : 1.0
  - Rdist : 0.5

Benchmarks

1. 일반 VQA: VQAv2, GQA(reasoning 요구), VizWiz(low vision quality), ScienceQA(logic and inference)

2. Scene Text VQA: TextVQA(text 인식 + context 이해)

3. Multimodal benchmarks: MME(broad multimodal eval), MMB(중국), MMBCN, POPE(hallucination measure), MMMU(Multi domain)

-> Avg7 = VQAv2, POPE, MMMU 제외

Experiment

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**

Table 1. **Benchmarked results with SoTA MLLMs**. Compared with counterparts, our LLaVA-KD achieves highly competitive results than current small-scale MLLM models. Optimal and sub-optimal results are in **bold** and underline. grey, orange and blue backgrounds represent ours baseline, the most direct MLLM distillation methods and our approach, respevtively. Avg_7: The average of the seven benchmarks for direct comparison with existing MLLM distillation methods, excluding VQAv2, POPE, MMMU. Avg_{10}: The average across all benchmarks for comprehensive comparison. †: reproduced results using the official code.

Ablation

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

Table 2. Ablation studies of different training stages.

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

Table 3. Ablation Study on MDist and RDist.

Training Stage	Response tokens	Prompt tokens	Visual tokens	Avg_{10}
DPT	✓	✗	✗	54.9
	✓	✓	✗	55.0
	✓	✗	✓	55.1
	✓	✓	✓	54.6
DFT	✓	✗	✗	57.2
	✓	✓	✗	56.9
	✓	✗	✓	57.7
	✓	✓	✓	57.1

Table 4. Ablation Study on Distillation Targets.

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**

Table 5. Comparison with Distillation Strategies in LLMs.

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**

Table 6. Comparison with distillation pipeline in LLaVA-MOD. MD and PD denote the Mimic Distillation and Preference Distillation.

LLM of the Teacher	LLM of the Student	Training Recipe	Avg_{10}
MLLaMA 2.7B	/	PT-SFT	55.2
/	MLLaMA 1.7B	PT-SFT	48.8
MLLaMA 2.7B	MLLaMA 1.7B	DPT-SFT	50.5
MLLaMA 2.7B	MLLaMA 1.7B	DPT-SFT-DFT	53.4

Table 7. Verification on MobileVLM.

LLM of the Teacher	LLM of the Student	Training Recipe	Avg_{10}
Qwen1.5-4B	/	PT-SFT	65.0
Qwen1.5-7B	/	PT-SFT	65.7
/	Qwen2.5-0.5B	PT-SFT	54.7
Qwen1.5-4B	Qwen1.5-0.5B	DPT-SFT-DFT	57.9
Qwen1.5-7B		DPT-SFT-DFT	57.4
Qwen2.5-3B	/	PT-SFT	66.2
Qwen2.5-7B	/	PT-SFT	69.3
/	Qwen2.5-0.5B	PT-SFT	56.5
Qwen2.5-3B	Qwen2.5-0.5B	DPT-SFT-DFT	58.9
Qwen2.5-7B		DPT-SFT-DFT	58.3

Table 8. Ablation study on teacher models with different sizes.

결론

장점

아키텍처 변경 없이 학습 패러다임만 개선

소형 MLLM 실사용 가능성 상승

한계

Teacher와 Student는 동일 vocab / LLM 계열 이어야 distillation이 원활

Teacher가 너무 크면 distillation 효율 떨어짐