Visual Representation Alignment for Multimodal Large Language Models

2025.10.21

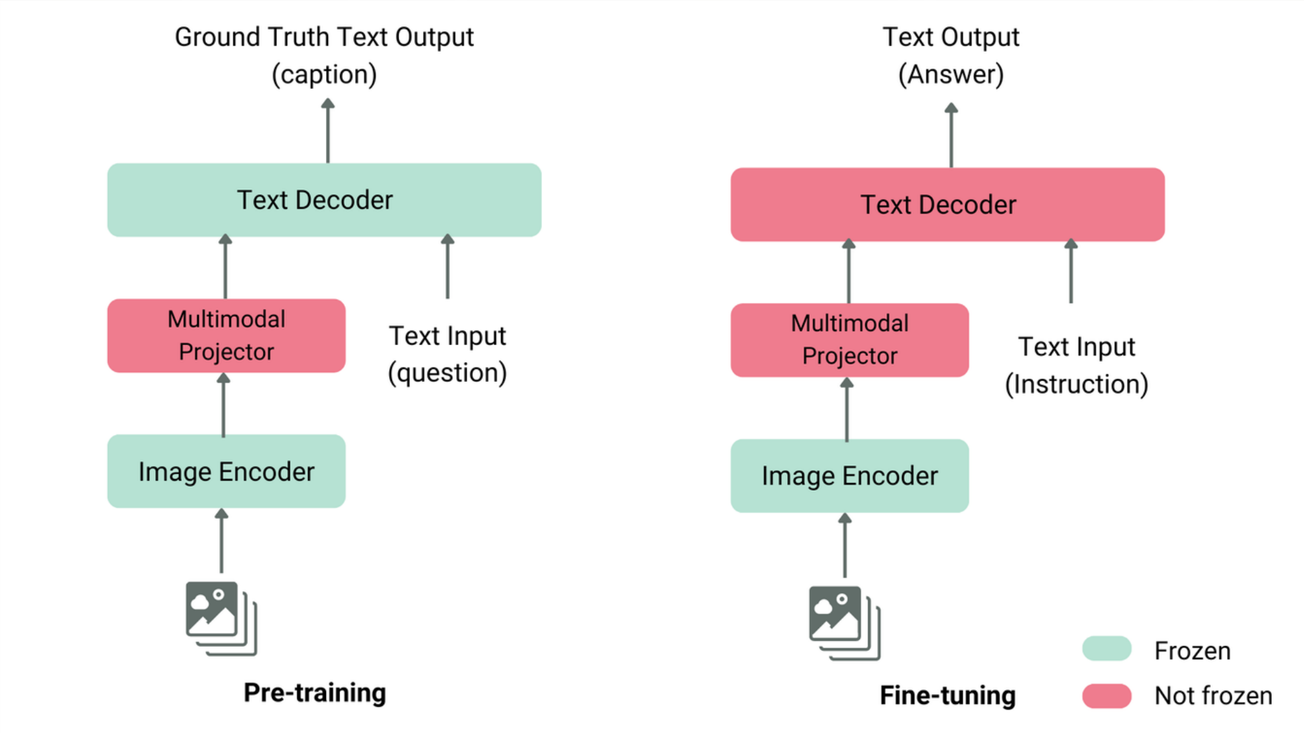
가짜연구소 박지예

https://arxiv.org/abs/2509.07979
https://huggingface.co/papers/2509.07979
https://github.com/cvlab-kaist/VIRAL

PseudoLab

Introduction

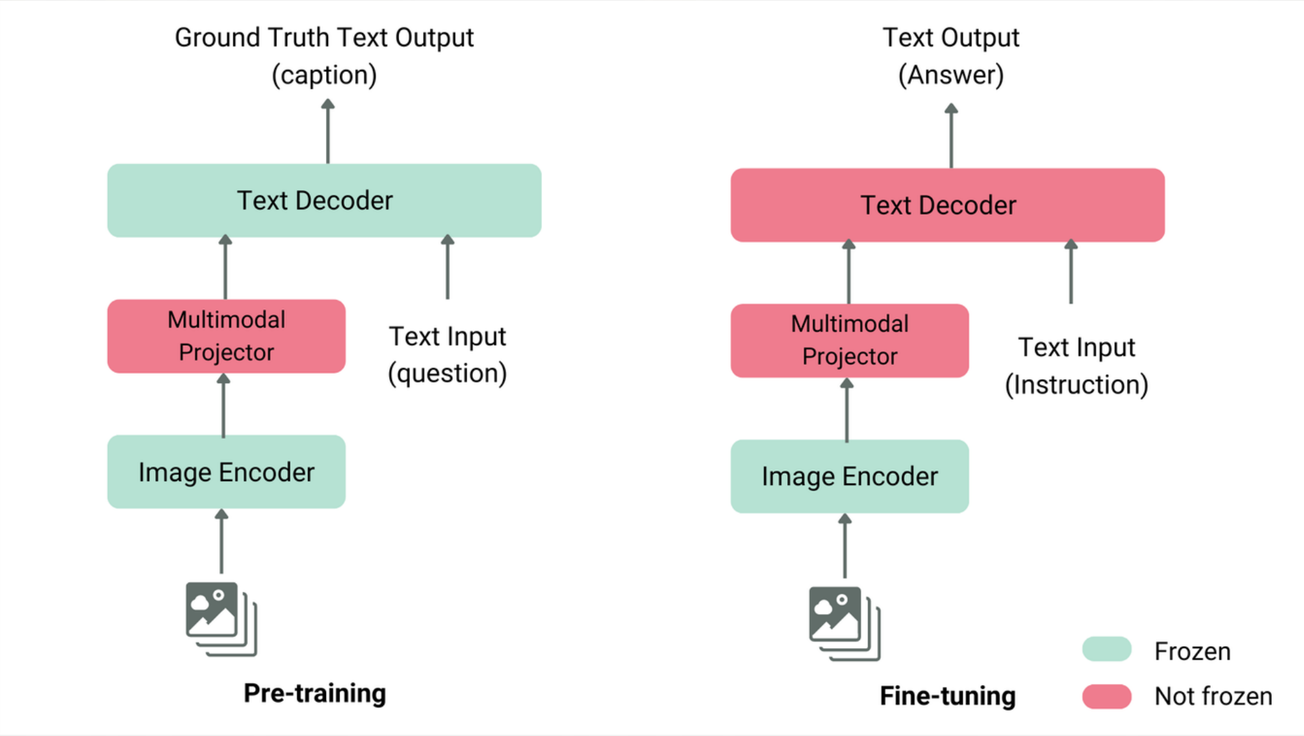
- Visual Instruction Tuning 방식을 적용한 MLLMs(Multimodal Large Language Models)이 좋은 성과를 보이고 있음
  - 사전 학습된 LLM + Vision Encoder + Projector로 구성
  - LLM이 시각적 문맥을 이해하고 다양한 과제를 수행 가능
- Projector: 시각 특징을 LLM이 이해할 수 있는 언어 임베딩으로 변환



Introduction

- 이전 연구에서 정밀한 시각 이해가 필요한 과제에서 한계 존재
  - 기존 연구들은 원인을 Vision Encoder나 Projector의 성능 부족으로 분석

  → 더 큰 모델 (DINOv2, CLIP 등) 이나 복잡한 프로젝터를 도입
- 하지만 이런 접근은 모델 크기 증가, 비용 상승, 확장성 저하 문제 초래



Introduction

- 대부분의 MLLM은 언어 생성 목적만으로 finetuning 되어, 시각적 정보는 **텍스트를 통해 간접적으로 학습**
  - 모델은 텍스트 예측에 필요한 정보만 남기고, 색상·위치·수량 등 세밀한 시각 정보는 잃게 됨
- 결과적으로, **vision encoder 표현과 MLLM 내부 시각 표현 간 불일치(misalignment) 발생**
- 이러한 문제를 해결하기 위해, 학습 중 MLLM의 내부 시각 표현을 vision encoder의 시각적 특징과 정렬
- 코사인 유사도 기반 정규화 loss로 정렬을 유도 → vision encoder의 세밀한 시각 정보가 MLLM 내부에서도 유지됨
- 더 강력한 Vision Foundation Model (VFM) 을 teacher로 사용하면 정렬 신호가 강화되어 시각 추론 성능이 향상됨

주요 기여점

- Revealed visual misalignment problem: 기존 Visual Instruction Tuning이 시각적 표현 불일치를 일으켜 시각 추론 능력 저하됨 확인
- Proposed VIRAL for visual alignment: VIRAL을 통해 encoder와 MLLM 내부 표현을 정렬, 세밀한 시각 정보 보존
- Achieved consistent performance gains: 여러 벤치마크에서 +9.4% 성능 향상, VIRAL의 효과 검증

Preliminaries

- Architecture of MLLMs
  - 입력 받은 이미지를, vision encoder(ψ)가 패치 단위로 특징 추출

$$z = V_\psi(I) \in R^{N \times D_z}$$ (N: 패치 개수, D_z: 각 시각적 feature 차원의 수)

  - Projector(φ)를 거쳐 $e_{img} \in \mathbb{R}^{N \times D}$ 로 변환
  - 텍스트 입력도 동일한 임베딩 공간으로 변환 후, [시각적 feature + 텍스트]가 하나의 시퀀스로 입력됨

  - 시각 정보를 조건으로, 순차적으로 텍스트 토큰이 생성되는 확률 추정

$$p_{\theta,\varphi}(e_{1:K}^{text} \mid e^{img}) = \prod_{i=1}^{K} p_{\theta,\varphi}(e_{1:K}^{text} \mid e^{img})$$

- MLLM은 이미지를 입력받아 다음 단어를 예측하는 Autoregressive 구조

Preliminaries

- Training Stages of MLLMs
  - Vision-Language Pretraining: 이미지-텍스트 캡션 데이터를 기반으로, '이미지 → 언어' 연관성 학습
  - Visual Instruction Tuning (VIT): 명령 기반 질의응답, 추론, 대화 학습

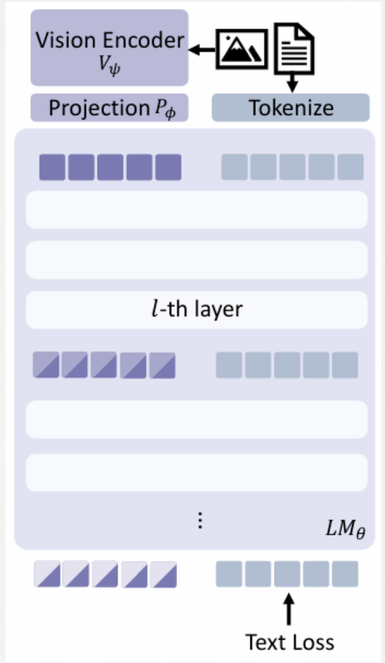
  - 텍스트 생성 중심의 Log-Likelihood Maximization 사용

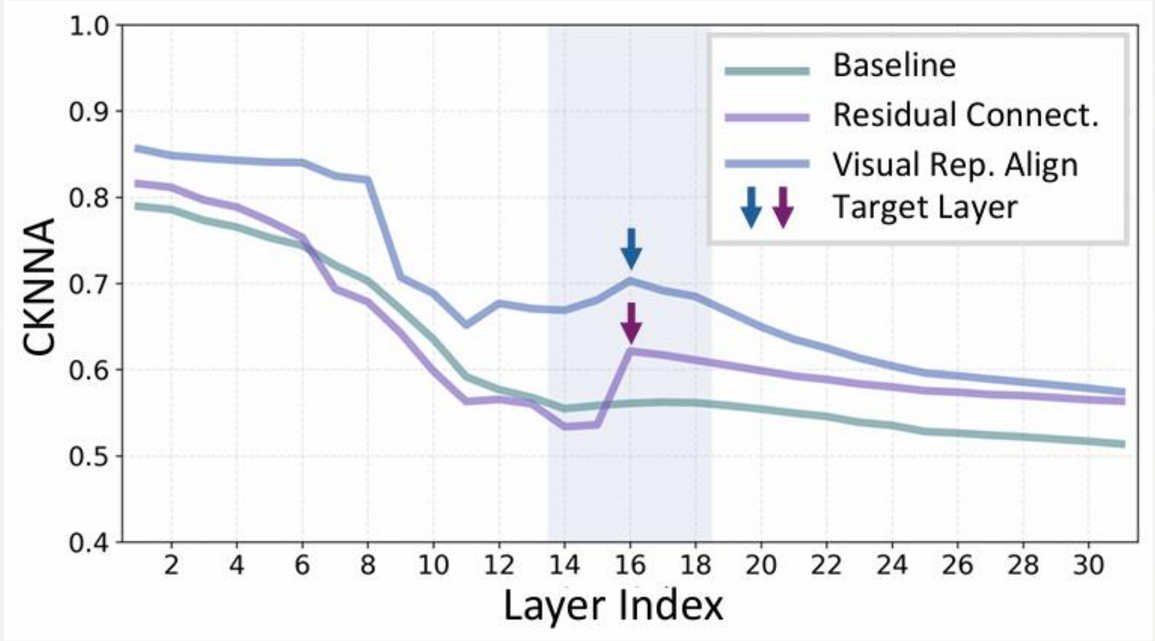
  → 시각 정보는 직접 학습되지 않고, 보조 입력 역할

$$L_{LM} = -\frac{1}{K}\sum_{i=1}^{K} log p_{\theta,\varphi}(e_1^{text} \mid e_{<i}^{text}, e^{img})$$

Methodology

- Do MLLMs undergo visual information loss?
  - MLLM은 이미지를 토큰으로 받아들이지만, 실제로는 텍스트 예측만 보고 배우기 때문에 시각 정보는 직접적으로 지시 받지 못함
  - 따라서 모델은 단어 예측에 필요한 남기고, 세밀한 시각 정보를 점차 버리게 됨
- LLaVA 모델과 CLIP을 이용해 CKNNA 유사도를 측정한 결과, 전체적으로 시각적 특징이 손실됨
- 중간층에서 시각적 근거를 기반으로 답을 형성하기 때문에, 일시적으로 시각 정보가 다시 활용됨



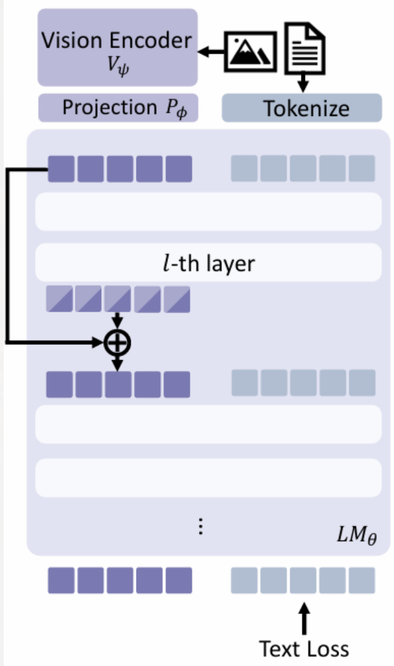


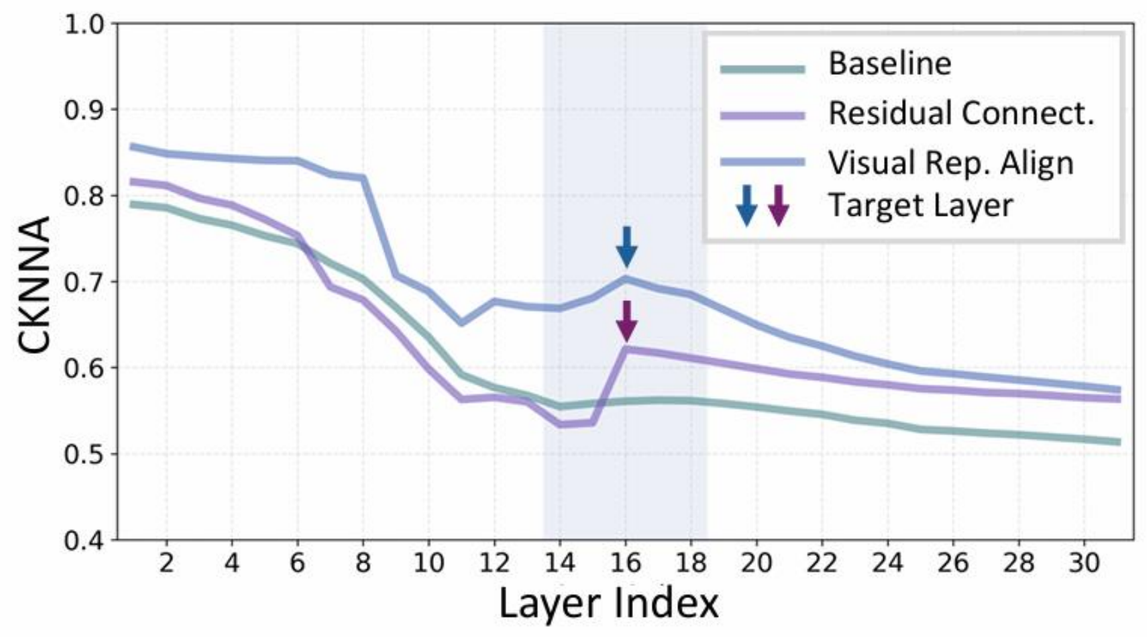
Methodology

- IS PRESERVING VISUAL INFORMATION BENEFICIAL?
  - 가설: MLLM이 중간층에서 시각적 특징을 더 보존하면, 모델의 시각 이해 능력이 향상될 것이다.
  - 방법: 모델의 중간층의 시각적 표현(e^l)에 vision encoder의 원래 시각 표현 z를 **residual connection**

$$e^l = e^l + P(z)$$

- Vision encoder의 feature z와의 유사도 증가
- 한계: projector (P)는 요약 과정을 거치기 때문에, 원래의 시각 표현을 완벽하게 전달하지 못함



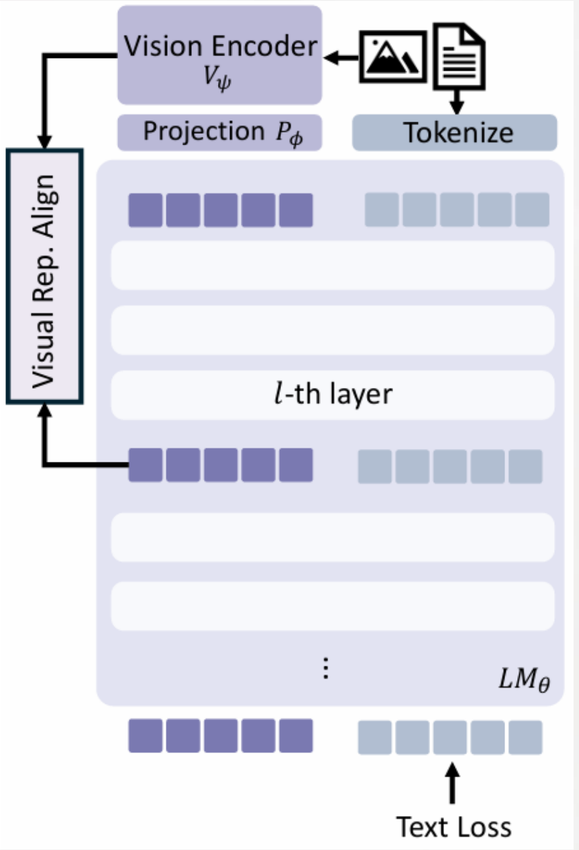


Methodology

- VISUAL REPRESENTATION ALIGNMENT FOR MLLMS
  - **MLLM의 중간층 시각 feature를 vision encoder의 feature space로 projection 시켜 정렬**
  - Vision encoder는 frozen (gradient 역전파 X)
- 새로운 학습 가능한 projection 모듈 도입
  - MLLM 내부 중간층 시각 표현을 vision encoder의 feature space로 mapping
  - 두 표현 간 코사인 유사도를 최대화

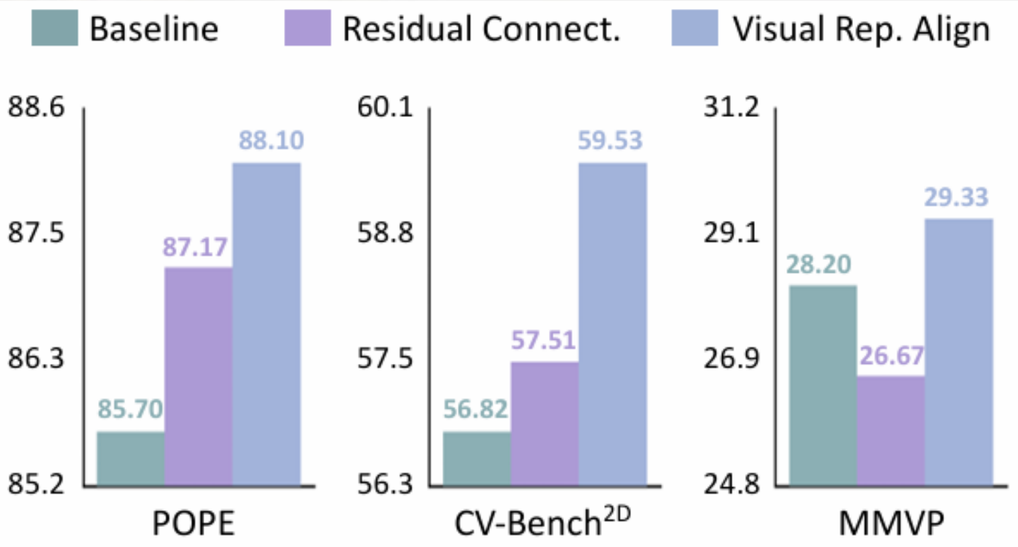
$$L_{VRA} = 1 - \cos(P(h^l), z)$$

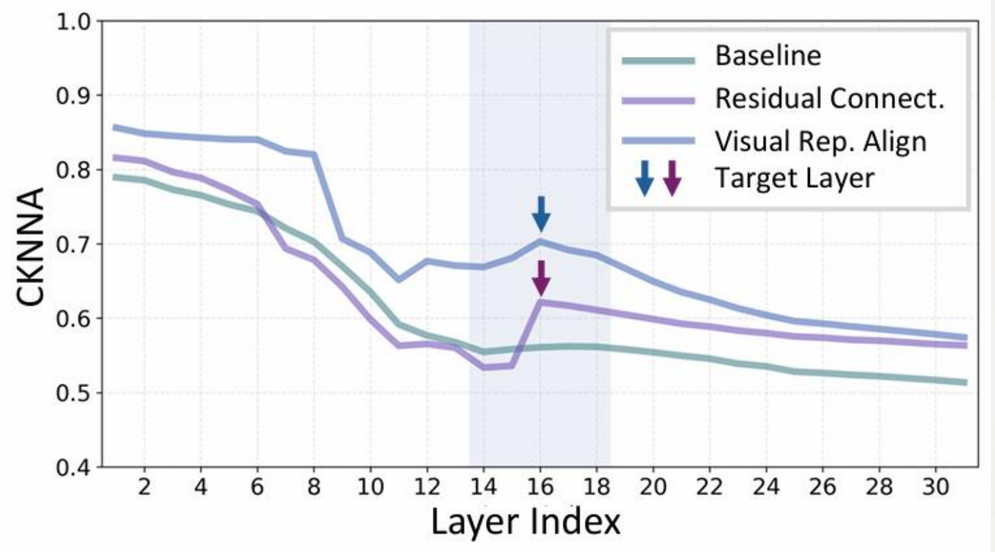
$$L_{total} = L_{LM} + \lambda L_{VRA}$$



Methodology

- Residual Connection보다 CKNNA 유사도와 벤치마크 성능이 더 높아짐
- 중간층의 표현이 vision encoder의 표현과 일관성 유지
- 세밀한 시각적 의미 보존 향상됨
- 한계
  - Encoder Dependence: encoder의 편향이 그대로 전이됨
  - Representation Constraint: CLIP의 표현이 충분히 풍부하지 않으면 성능이 정체됨
  - Alignment Target Problem: CLIP에서의 표현과 정렬시키는 것으로 충분하지 않을 수 있음



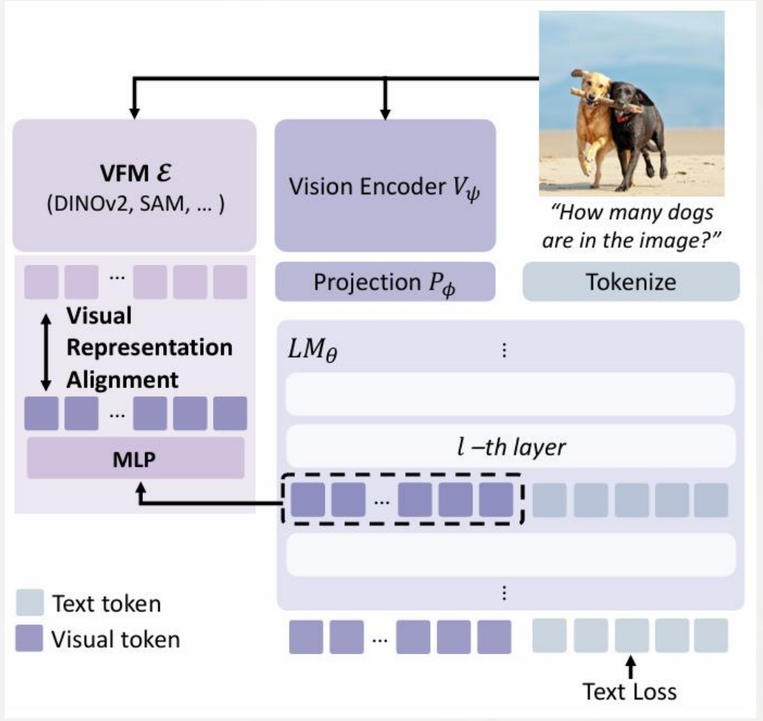


Methodology

- From encoder features to other VFMs
  - CLIP은 텍스트-이미지 정렬에는 강하지만, 세부적 시각 정보 보존에는 한계가 있음
  - CLIP 대신, Vision Foundation Model(VFM)이 학습한 feature space에 직접 mapping 시킴
- VIRAL with Vision Foundation Models

$$L_{VRA} = 1 - \cos\left(P_\pi\left(h^l\right), v\right), \qquad v = E(I)$$

  - MLLM의 중간층 시각 표현이 VFM feature와 유사하도록 학습
- VIRAL이 더 구조적이고 세밀한 시각 의미 보존
- 모델이 텍스트 중심 이해에서, 시각적 이해로 확장됨



Experiments

- Experimental Setup
  - Baseline MLLM: LLaVA-1.5
  - LLM (Backbone): Vicuna-1.5-7B, Qwen2.5-7B, Vicuna-1.5-13B
  - Vision Encoder: CLIP, SigLIP-v2
  - Alignment Module P_π: 3-layer MLP + SiLU
  - Alignment Target $\mathcal{E}(\cdot)$: DINOv2
- Evaluation
  - Vision-centric (정밀 시각 추론/카운팅): CV-Bench, What's Up, MMVP
  - Hallucination detection (시각 환각 억제): POPE, MMStar
  - General multimodal: MME

Experiments

- VIRAL (L_{VRA}) 적용 효과

Language Model	Vision Encoder	$\mathcal{L}_{\text{VRA}}$	CV-Bench^{2D}	MMVP	What's Up	POPE	MMStar	MME
Vicuna-1.5-7B	CLIP	✗	56.82%	28.20%	40.13%	85.70%	**33.93%**	1650.21
		✓	**59.67%** (+2.85)	**33.33%** (+5.13)	**48.55%** (+8.42)	**88.32%** (+2.62)	**33.93%** (±0.00)	**1694.52** (+44.31)
	SigLIPv2	✗	58.90%	28.22%	40.90%	90.13%	36.53%	1738.96
		✓	**62.66%** (+3.76)	**33.11%** (+4.89)	**44.40%** (+3.50)	**90.77%** (+0.64)	**37.20%** (+0.67)	**1835.62** (+96.66)
Qwen2.5-7B	CLIP	✗	58.97%	33.47%	59.08%	**85.88%**	39.20%	1743.56
		✓	**60.50%** (+1.53)	**36.07%** (+2.60)	**63.57%** (+4.49)	84.92% (-0.96)	**39.67%** (+0.47)	**1765.65** (+22.09)
Vicuna-1.5-13B	CLIP	✗	57.51%	32.30%	44.44%	87.12%	34.47%	1599.04
		✓	**58.97%** (+1.46)	**37.80%** (+5.50)	**62.26%** (+17.82)	**87.79%** (+0.67)	**37.00%** (+2.53)	**1636.62** (+37.58)

  - VIRAL 추가만으로 모든 조합에서 일관된 성능 향상 보임
  - 강한 vision encoder (SigLIP-v2)를 쓰더라도 추가적인 성능 향상이 지속됨 (**정렬 자체에서 효과가 나타남**)

Experiments

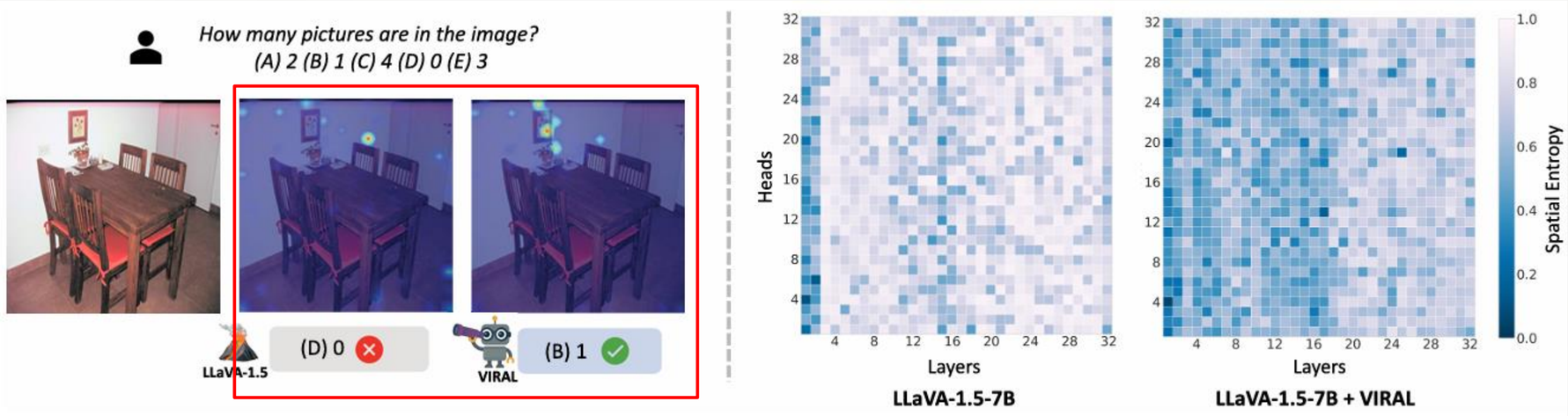
- Target Layers
  - 각 layer에 정렬 적용 실험: 매 4번째 층마다 평가 진행

VFM	Layer Index	CV-Bench^{2D}	MMVP	What's Up	POPE	MME
Baseline		56.82%	28.20%	40.13%	85.70%	1650.21
Ablation studies on different VFMs						
DINOv2	16	**59.67%**	**33.33%**	48.55%	88.32%	**1694.52**
CLIP	16	59.53%	29.33%	44.50%	88.10%	1548.49
SAM	16	57.58%	30.27%	**49.84%**	88.34%	1648.77
DAv2	16	58.55%	28.67%	47.29%	**88.70%**	1682.42
RADIO	16	57.59%	31.80%	47.35%	88.52%	1692.94
Ablation studies on different target layers						
DINOv2	4	58.55%	30.67%	45.05%	87.68%	1720.36
DINOv2	8	58.28%	27.70%	48.32%	88.43%	1662.67
DINOv2	12	57.77%	28.59%	48.19%	88.27%	1648.88
DINOv2	16	**59.67%**	**33.33%**	**48.55%**	88.32%	1694.52
DINOv2	20	55.22%	27.41%	48.04%	88.39%	1705.97
DINOv2	24	55.77%	27.48%	47.99%	88.10%	1740.55
DINOv2	28	54.87%	27.19%	47.82%	**88.56%**	**1755.86**
DINOv2	32	56.12%	26.52%	47.60%	87.32%	1678.69

  - 16번째 layer(중간층)에서 가장 안정적인 성능 향상 → 세밀한 시각 정보가 중간층에서 가장 잘 유지됨

Experiments

- Attention Analysis
  - VIRAL이 텍스트-이미지 attention map을 어떻게 변화시키는지 분석
  - 시각적으로 의미 있는 영역에 집중하는지 평가



  - Baseline: attention이 불분명하고 분산됨
  - VIRAL 적용: attention이 시각적으로 의미 있는 객체 영역에 집중됨, 불필요한 영역 무시

Experiments

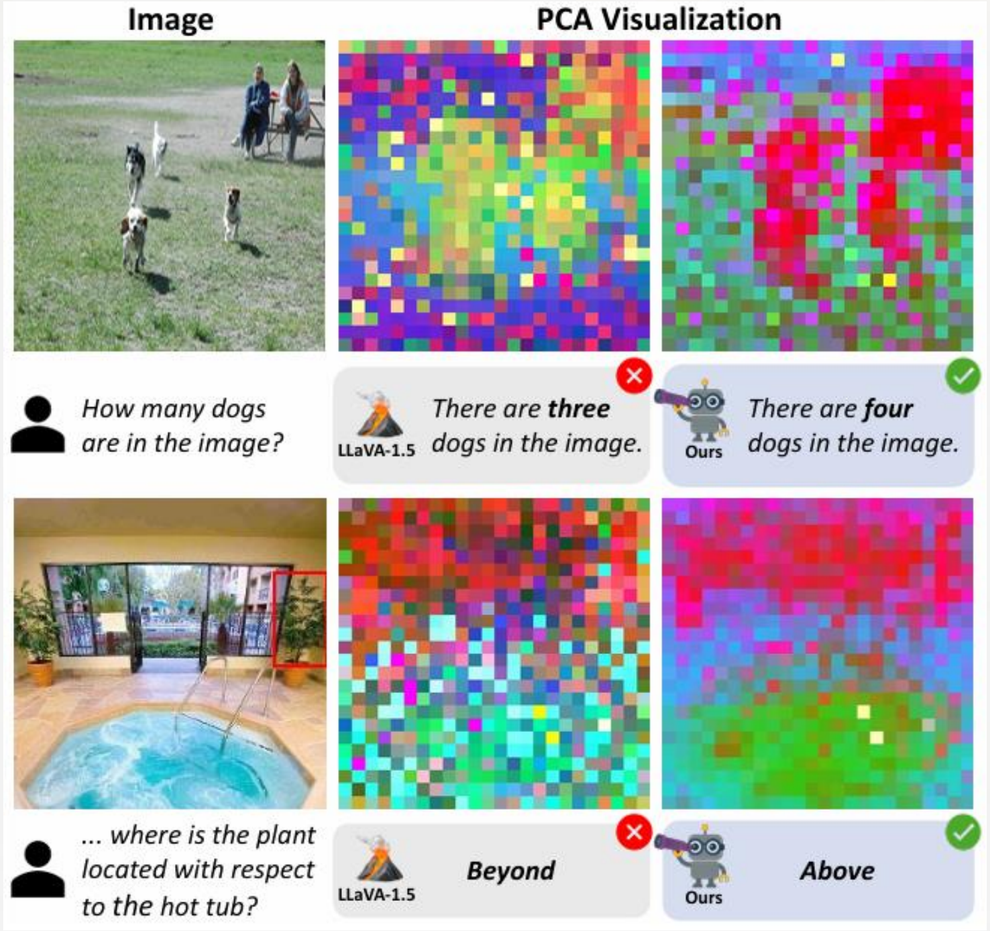
- Robustness Analysis
  - 시각적 토큰 순서를 임의로 섞었을 때, 모델이 구조적 정보에 대해 민감한 정보 평가
  - 시각적으로 의미 있는 영역에 집중하는지 평가

Vision Enc.	$\mathcal{L}_{\text{VRA}}$	original	patch shuffle	Δ
CLIP	✗	400	374	−26 (6.5%)
	✓	414	360	**−54 (13.0%)**
SigLIPv2	✗	374	353	−21 (5.6%)
	✓	436	353	**−83 (19.0%)**

  - VIRAL 모델이 성능 하락폭이 더 큼
    → 모델이 실제 공간적 정보를 더 인식하고 활용하기 때문에, 공간 관계에 더 민감함
  - VIRAL이 단순히 평균 성능을 향상시킬 뿐만 아니라, 세밀한 공간적 추론 능력도 강화 시킴

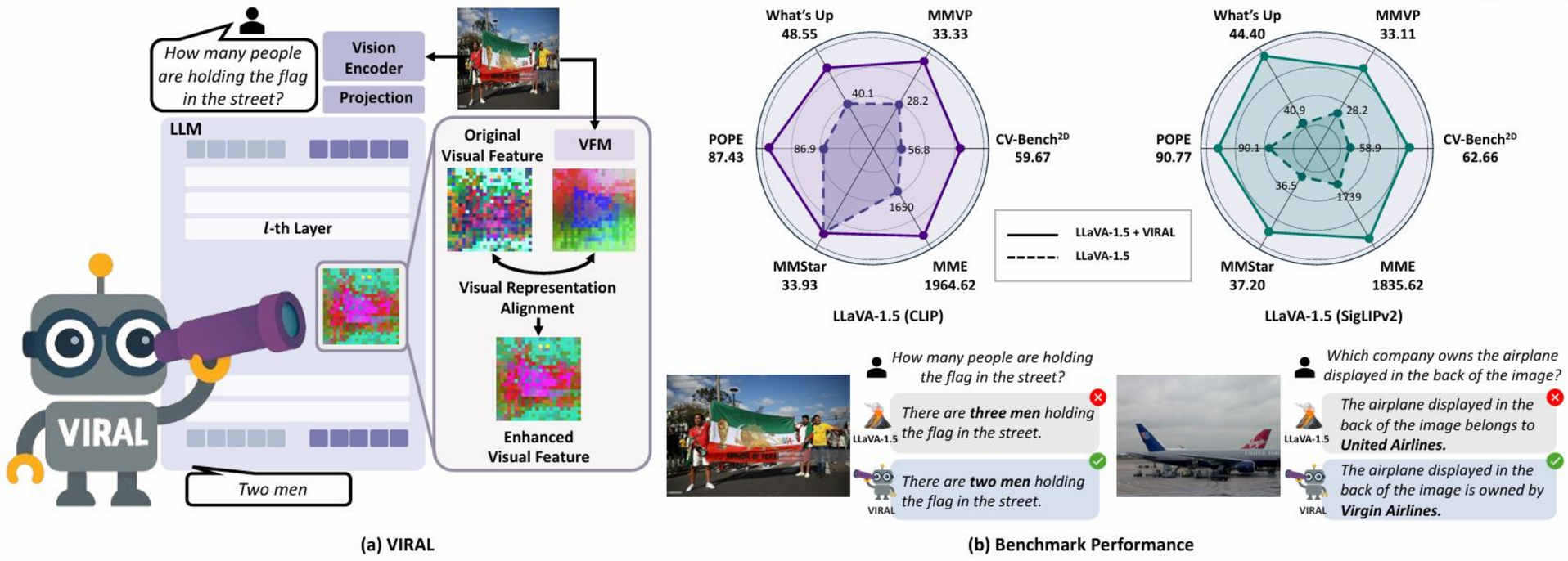
Experiments

- Qualitative Results
  - 구조적으로 더 의미론적으로 일관된 feature embedding 형성



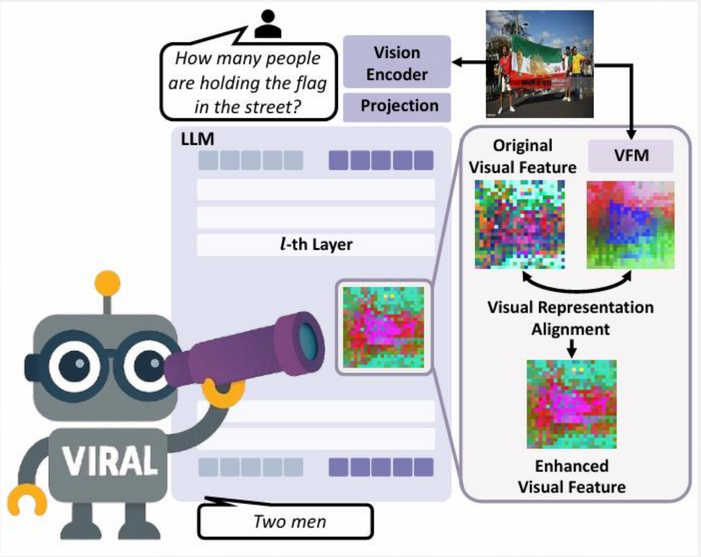
Conclusion

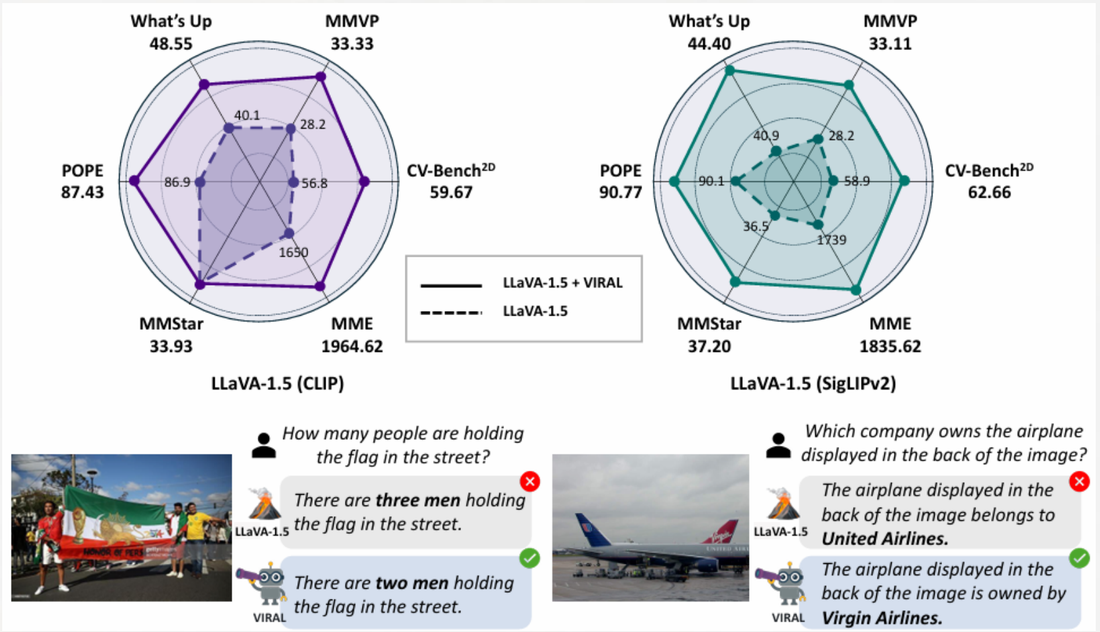
- VIRAL: MLLM의 내부적 시각 표현을 사전 학습된 VFM 표현과 정렬하는 효과적인 정규화 전략
- 효과
  - Fine-grained visual semantics 보존
  - Spatial reasoning과 object grounding 강화
  - Attention 집중도 향상



Visual Representation Alignment for Multimodal Large Language Models

[VIRAL]





[Benchmark Performance]

- Motivation
  - 기존 MLLM은 이미지-텍스트를 함께 학습하지만, 텍스트 예측을 중심으로 학습을 진행하기 때문에 시각 정보가 사라짐
  - Fine-grained Spatial Reasoning에서 성능 저하 발생
- Methodology: VIRAL(VIsual Representation ALignment)
  - MLLM의 중간층 시각 표현을 사전학습된 VFM의 특징과 정렬시키는 정규화 기법
  - 기존 LM loss에 Alignment loss(L_{VRA}) 추가
  - Spatial reasoning과 object grounding 강화
  - Attention 집중도 향상
- Experimental Results
  - 평균 +9.4% 성능 향상 확인, 특히 시각 중심 과제(vision-centric tasks)에서 두드러짐
  - 세밀한 시각 추론 능력 향상, 객체 인식과 시각 근거(object grounding) 강화 입증