DMQA Open Seminar

Visual Reasoning in Multimodal Large Language Models

2026. 10. 30

김성수

Data Mining and Quality Analytics Lab

고려대학교
KOREA UNIVERSITY

발표자 소개



❖ 김성수 (Sungsu Kim)

- 경희대학교 산업경영공학과 학부 졸업 (2022.02)
- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석박통합 과정 (2022.03 ~ Present)

❖ Research Interest

- Computer Vision
- Vision-Language Model

❖ Contact

- 2022020650@korea.ac.kr

Introduction

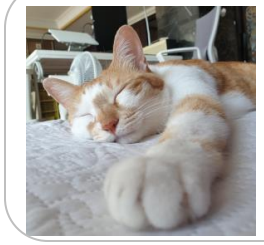
❖ Multimodal Large Language Model

- **Large Language Model (LLM):** 챗봇처럼 다양한 인간의 텍스트 질문에 대답할 수 있는 모델
- **Multimodal Large Language Model (MLLM):** 이미지를 포함한 소통이 가능한 LLM

LLM: 텍스트 → 텍스트

Q. 고양이는 왜 잠을 많이 잘까? → LLM → A. 에너지를 아끼기 위해 하루의 대부분을 잠으로 보냅니다.

MLLM: 이미지 + 텍스트 → 텍스트



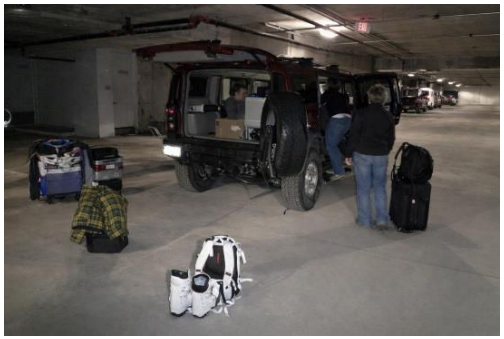
Q. 이 고양이는 지금 뭘 하고 있어? → MLLM → A. 침대 위에 몸을 쭉 뻗고 편하게 잠을 자고 있습니다.

고려대학교 KOREA UNIVERSITY

Introduction

❖ **Computer Vision Task w/ MLLM → Visual Reasoning**

- 이미지를 고려한 인간 질문에 답변 가능
- 특정 Vision Task에 한정되지 않고, 다양한 Task 수행 가능
- 심지어, 해당 출력에 대한 근거 등 고도화된 추론까지 가능



입력 이미지



MLLM

Q. 이 장소는 어디야?
A. 실내 주차장입니다.

Q. 자동차는 무슨 색깔이야?
A. 회색 바닥보다 더 어둡습니다. 따라서 검정색 입니다.

고려대학교 KOREA UNIVERSITY

Introduction

❖ MLLM이 모든 Visual Task를 잘 수행하고 있을까?

- 단순한 종류의 Visual Task에는 우수 (ex. 이미지 분류)
- 이미지 기반 질의 응답 (Visual Q&A, VQA) 등 고도의 생각이 필요한 Task에서는 여전히 저조
- 이미지를 활용하나, 대부분의 추론을 언어 공간에서 수행 → 세밀한 시각 정보 부족

단순 Visual Task (ex. 이미지 분류)



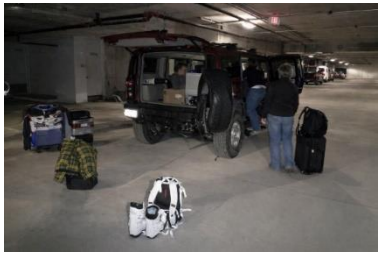
MLLM

A. 고양이

정답

Q. 이 동물은 뭐야?

고도의 추론 Task (ex. VQA)



MLLM

A. 2개

오답

Q. 바닥에 놓인 가방은 모두 몇 개야?

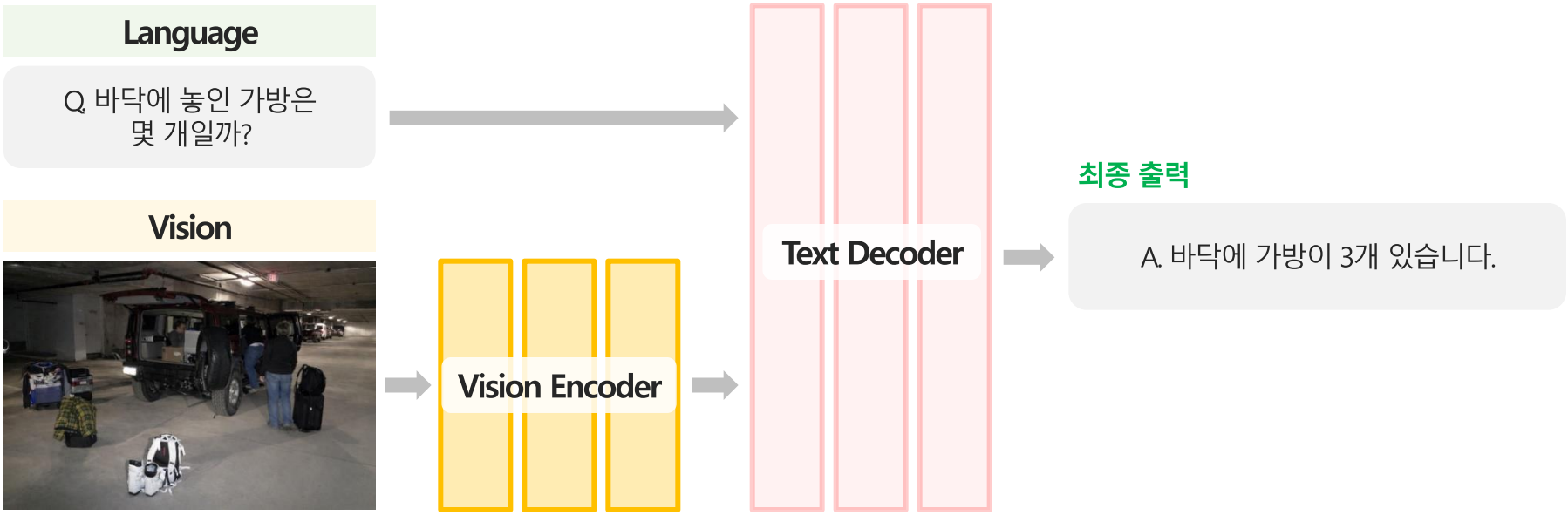
→ 작은 물체·개수 등 세밀한 시각 정보를 놓침

고려대학교 KOREA UNIVERSITY

Introduction

❖ **MLLM이 모든 Visual Task를 잘 수행하고 있을까?**

- 단순한 종류의 Visual Task에는 우수 (ex. 이미지 분류)
- 이미지 기반 질의 응답 (Visual Q&A, VQA) 등 고도의 생각이 필요한 Task에서는 여전히 저조
- 이미지를 활용하나, 대부분의 추론을 언어 공간에서 수행 → 세밀한 시각 정보 부족



어떻게 Visual Reasoning에서 MLLM을 잘 활용할 수 있을까?

고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

❖ CCoT (2024/CVPR – UC Berkeley)

- 인용수: 470회
- 텍스트 프롬프트를 개선하여 MLLM의 Visual Reasoning 성능 개선

Compositional Chain-of-Thought Prompting for Large Multimodal Models

Chancharik Mitra Brandon Huang Trevor Darrell Roei Herzig
University of California, Berkeley

Abstract

The combination of strong visual backbones and Large Language Model (LLM) reasoning has led to Large Multimodal Models (LMMs) becoming the current standard for a wide range of vision and language (VL) tasks. However, recent research has shown that even the most advanced LMMs still struggle to capture aspects of compositional visual reasoning, such as attributes and relationships between objects. One solution is to utilize scene graphs (SGs)—a formalization of objects and their relations and attributes that has been extensively used as a bridge between the visual and textual domains. Yet, scene graph data requires scene graph annotations, which are expensive to collect and thus not easily scalable. Moreover, finetuning an LMM based on SG data can lead to catastrophic forgetting of the pretraining objective. To overcome this, inspired by chain-of-thought methods, we propose Compositional Chain-of-Thought (CCoT), a novel zero-shot Chain-of-Thought prompting method that utilizes SG representations in order to extract compositional knowledge from an LMM. Specifically, we first generate an SG using the LMM, and then use that SG in the prompt to produce a response. Through extensive experiments, we find that the proposed CCoT approach not only improves LMM performance on several vision and language (VL)

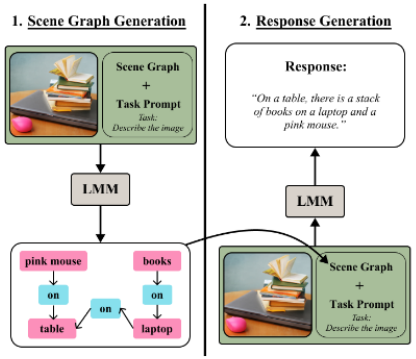


Figure 1. **A high-level overview of our Compositional Chain-of-Thought (CCoT) approach.** Our CCoT method consists of a two-step prompting process: (1) First, the LMM is prompted to generate a scene graph relevant to the image and task prompt, such as the task in the figure "Describe the image". (2) Following this, the LMM is prompted with the generated scene graph, the image, and the task prompt as context for responding in a way that incorporates the compositional information in the scene graph to provide a correct description of the complex scene.

[1] Mitra, C., Huang, B., Darrell, T., et al. (2024). Compositional Chain-of-Thought Prompting for Large Multimodal Models. In CVPR.

고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

❖ CCoT (2024/CVPR) – 왜 MLLM이 Visual Reasoning이 저조할까?

- **문제상황:** 이미지 내 객체 종류는 잘 파악, 그러나 세부적인 관계는 파악하지 못함



입력 이미지

객체 종류

노트북

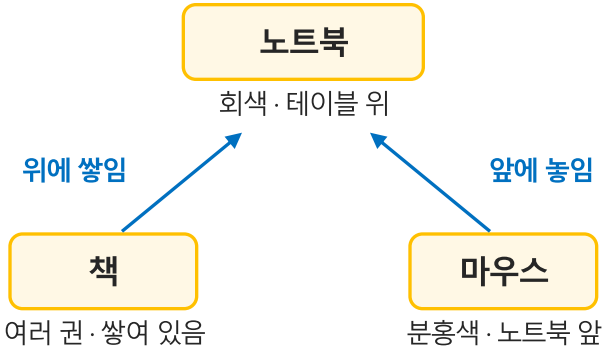
책

마우스

테이블

→ 객체 간 관계·속성 정보 부족

객체 + 속성 + 관계



→ 구조화된 시각 정보로 추론 가능

고려대학교 KOREA UNIVERSITY

Algorithm

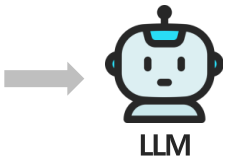
- (1) 텍스트 프롬프트 기반 접근

❖ CCoT (2024/CVPR) – 어떻게 Visual Reasoning을 개선할 수 있을까?

- **Chain-of-Thought (CoT):** 텍스트 프롬프트에 단계적으로 생각하는 방법을 명시
  → MLLM이 단계적으로 생각할 수 있도록 유도
- **연구동향:** LLM에서는 CoT를 잘 작성하는 것만으로도 성능이 크게 향상
- **해결방향:** MLLM도 CoT를 잘 작성하면 성능이 개선되지 않을까? → "텍스트 프롬프트 개선"

일반 프롬프트

Q. 사과 5개 중 2개를 먹고, 3개를 더 샀다.
지금 사과는 몇 개일까?

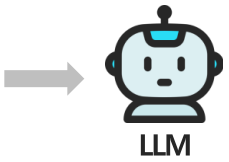


A. 8개

오답

CoT 프롬프트

Q. 사과 5개 중 2개를 먹고, 3개를 더 샀다.
지금 사과는 몇 개일까?
Let's think step by step.
① 없어진 사과를 먼저 계산해라
② 새로 생긴 사과를 계산해라



① 5개 중 2개를 먹음 → 3개
② 3개를 더 삼 → 3 + 3 = 6개
A. 6개

정답

고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

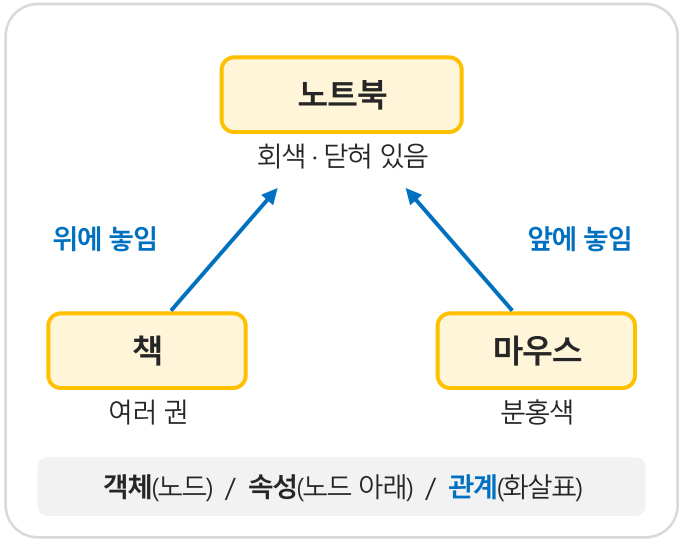
❖ CCoT (2024/CVPR) – 어떤 식으로 CoT를 구성할까?

- **CoT 작성 방향:** MLLM이 부족한 객체 관계성에 대한 Compositional CoT (CCoT)구성
- **CCoT:** Scene Graph 개념을 텍스트화 해서 CoT 작성
- **Scene Graph:** 객체 종류 / 해당 객체의 속성 / 각 객체의 관계를 표현하는 Graph



입력 이미지

Scene Graph



텍스트화

CCoT: 텍스트로 쓴 Scene Graph

```
{
  "objects": [
     "books", "laptop", "mouse" ],
  "attributes": [
     {"laptop": "gray"},
     {"mouse": "pink"} ],
  "relationships": [
     {"books": {"on": "laptop"}} ]
}
```

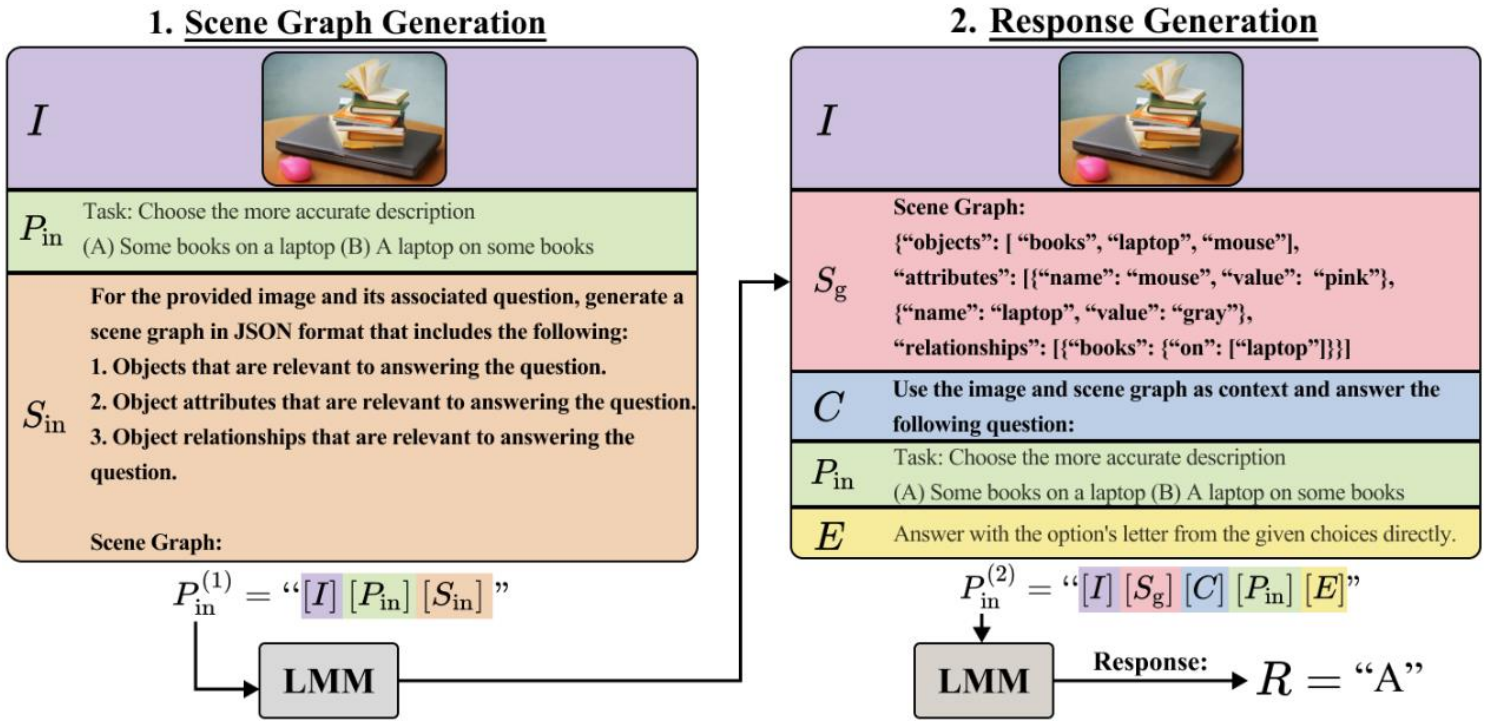

고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

❖ **CCoT (2024/CVPR) – CCoT 방법론은 어떻게 구성될까? (1) Scene Graph Generation**

- **입력:** 이미지 + 질문 + Scene Graph 생성 지시문
- **출력:** 질문과 관련된 객체 / 속성 / 관계를 JSON 형식으로 정리



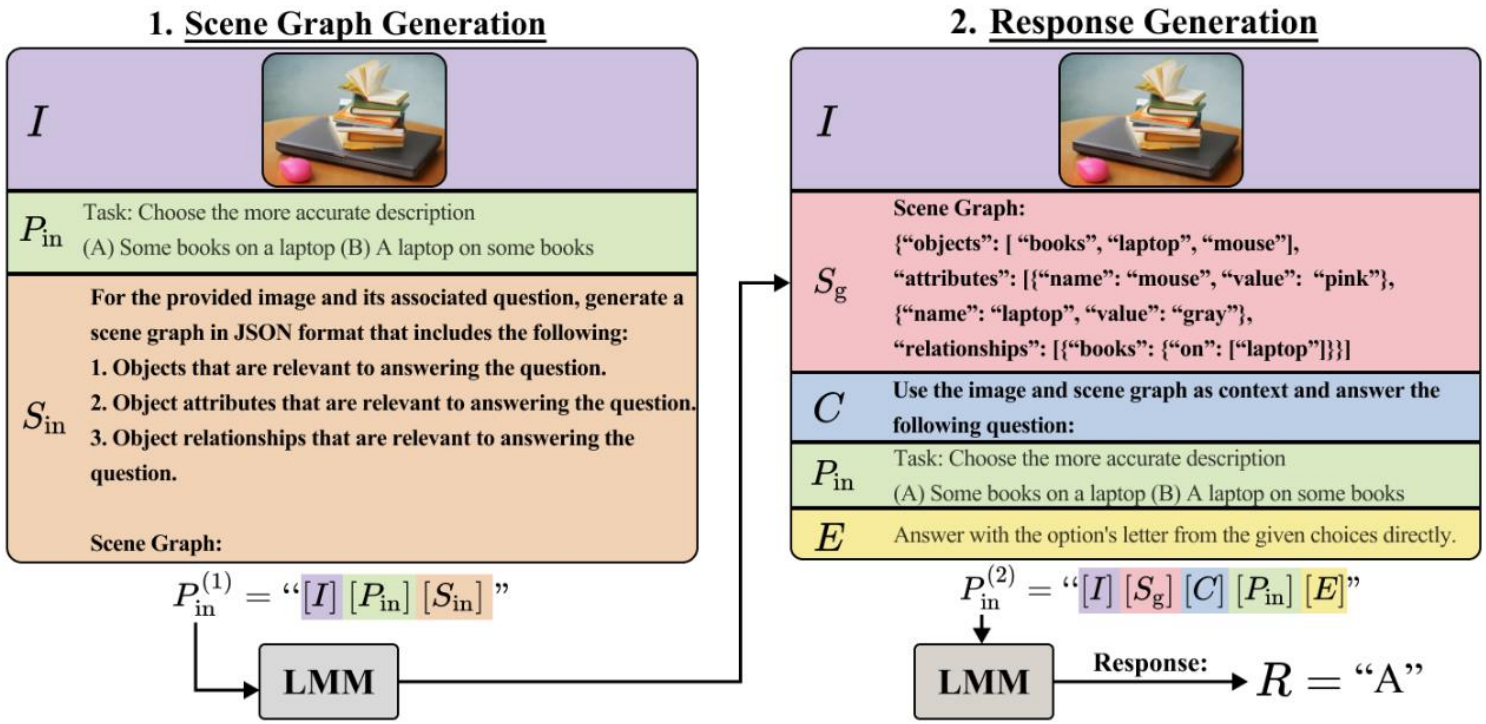
고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

❖ **CCoT (2024/CVPR) – CCoT 방법론은 어떻게 구성될까? (2) Response Generation**

- **입력:** 이미지 + 생성된 Scene Graph + 질문 + 답변 형식 지시문
- **출력:** Scene Graph를 근거로 추론한 최종 답변 (ex. "A")



고려대학교 KOREA UNIVERSITY

Algorithm

- (1) 텍스트 프롬프트 기반 접근

❖ CCoT (2024/CVPR) – 실험결과

- 객체 관계 관련된 벤치마크 뿐만 아니라, 세밀한 이해가 필요한 여러 Visual Task에서 효과적
- MLLM 모델에 관계없이 모두 효과적

	Multimodal Benchmarks			VL Compositional Benchmarks			
Model	SEED-I	MMBench	LLaVA-W	Wino-Text	Wino-Image	Wino-Group	WHOOPS! VQA BEM
CLIP	-	-	-	30.7	10.5	8.0	-
BLIP	-	-	-	39.0	19.2	15.0	39.0
BLIP2	46.4	-	-	42.0	23.8	19.0	55.0
SGVL†	-	-	-	42.8†	28.5†	23.3†	-
mPlug-OWL2	57.8	64.5	-	-	-	-	-
QwenVL-Chat	58.2	61.2	-	-	-	-	-
InstructBLIP-13B	48.2	36.0	47.2	12.8	13.3	4.5	48.3
InstructBLIP-13B-ZS-CoT	37.6	25.3	45.4	15.8	14.8	6.0	43.36
InstructBLIP-13B-CCoT	56.9 (+8.7)	40.3 (+4.3)	47.9 (+0.7)	26.0 (+13.2)	27.0 (+13.7)	11.5 (+7.0)	62.9 (+14.6)
LLaVA-1.5-13B	68.2	67.0	73.5	33.5	35.0	17.3	47.3
LLaVA-1.5-13B-ZS-CoT	66.7	66.0	68.5	36.8	35.0	19.8	46.6
LLaVA-1.5-13B-CCoT	69.7 (+1.5)	70.7 (+3.7)	74.9 (+1.4)	39.8 (+6.3)	37.3 (+2.3)	22.3 (+5.0)	61.2 (+13.9)
Sphinx	71.6	65.9	70.0	29.0	29.0	16.3	50.0
Sphinx-ZS-CoT	70.3	65.5	69.8	36.0	38.5	21.5	60.4
Sphinx-CCoT	74.2 (+2.6)	68.3 (+2.4)	71.0 (+1.0)	36.5 (+7.5)	36.3 (+7.3)	22.5 (+6.2)	61.9 (+11.9)
GPT4V	69.1	75.5	88.2	60.3	45.3	33.5	64.8
GPT4V-ZS-CoT	72.5	74.8	88.8	63.3	52.5	41.0	65.5
GPT4V-CCoT	74.0 (+4.9)	76.3 (+0.8)	91.2 (+2.0)	64.0 (+3.7)	54.5 (+9.2)	43.3 (+9.8)	67.8 (+3.0)

고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR – USC)

- 인용수: 214회
- 입력 이미지를 개선하여 MLLM의 Visual Reasoning 성능 개선

MLLMs Know Where to Look:
Training-free Perception of
Small Visual Details with Multimodal LLMs

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski

University of Southern California, USA
Vrije Universiteit Amsterdam, The Netherlands

Abstract

Multimodal Large Language Models (MLLMs) have experienced rapid progress in visual recognition tasks in recent years. Given their potential integration into many critical applications, it is important to understand the limitations of their visual perception. In this work, we study whether MLLMs can perceive small visual details as effectively as large ones when answering questions about images. We observe that their performance is very sensitive to the size of the visual subject of the question, and further show that this effect is in fact causal by conducting an intervention study. Next, we study the attention patterns of MLLMs when answering visual questions, and intriguingly find that they consistently know where to look, even when they provide the wrong answer. Based on these findings, we then propose training-free visual intervention methods that leverage the internal knowledge of any MLLM itself, in the form of attention and gradient maps, to enhance its perception of small visual details. We evaluate our proposed methods on two widely-used MLLMs and seven visual question answering benchmarks and show that they can significantly improve MLLMs' accuracy *without requiring any training*. Our results elucidate the risk of applying MLLMs to visual recognition tasks concerning small details and indicate that visual intervention using the model's internal state is a promising direction to mitigate this risk.[1]

[2] Zhang, J., Khayatkhoei, M., Chhikara, P., et al. (2025). MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In ICLR.

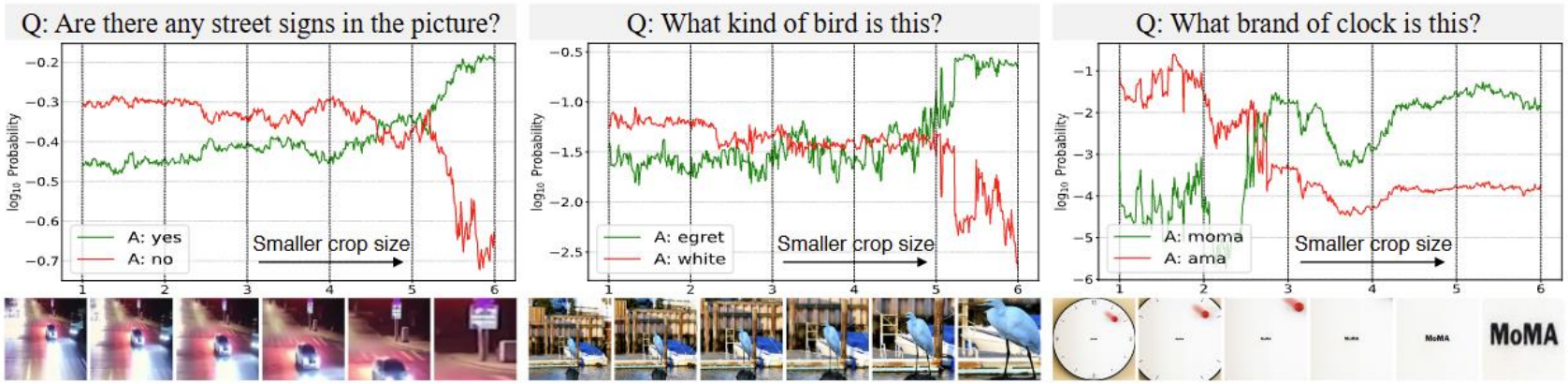
고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ **ViCrop (2025/ICLR) – 왜 MLLM이 Visual Reasoning에서 저조할까?**

- **문제상황:** MLLM은 타겟 대상이 작으면 잘 파악하지 못한다.
- **간단한 분석:** 타겟 대상을 Crop 후 확대하여 입력 → 성능이 개선되는 것을 확인



고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR) – 왜 MLLM이 Visual Reasoning에서 저조할까?

- **문제상황:** MLLM은 타겟 대상이 작으면 잘 파악하지 못한다.
- **심층분석:** 다양한 모델에서 객체가 작을수록 Crop이 효과적
  - ➢ Human-crop: 객체의 GT BBox를 기준으로 Crop 후 원본 이미지 크기만큼 확대

① 객체 상대 크기 (S)

S = 정답 BBox 면적 ÷ 전체 이미지 면적



small
S < 0.005

medium
0.005 ≤ S < 0.05

large
S ≥ 0.05

② Human-CROP 정답 BBox 기준 정사각형 Crop



정답 BBox → Crop & Resize → **MLLM**

Model	Method	Answer Bbox Size (S)		
		small	medium	large
BLIP-2 (FlanT5$_{XL}$)	no cropping	12.13	19.57	36.32
	human-CROP	55.76	52.02	45.73
InstructBLIP (Vicuna-7B)	no cropping	21.79	30.58	45.30
	human-CROP	69.60	61.56	53.39
LLaVA-1.5 (Vicuna-7B)	no cropping	39.38	47.74	50.65
	human-CROP	69.95	65.36	56.96
Qwen-VL (Qwen-7B)	no cropping	56.42	65.09	68.60
	human-CROP	70.35	75.49	71.05
GPT-4o	no cropping	65.76	72.81	69.17
	human-CROP	75.63	81.32	71.72

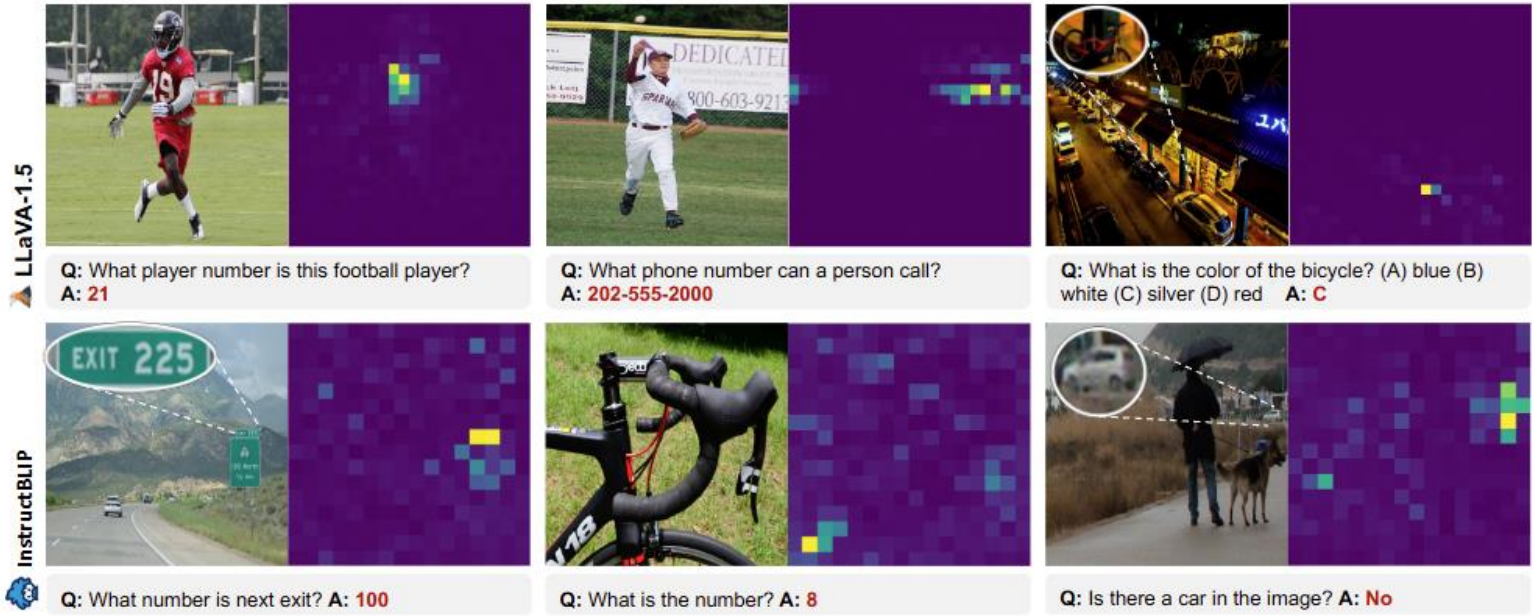
고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR) – MLLM은 어디를 봐야 할지 알까?

- **분석:** 답변 시 MLLM의 Attention이 정답 영역에 얼마나 집중되는지 측정 (Attention Map)
- **결과:** 오답일 때도 정답 영역에 높은 Attention → 위치는 안다, 그러나 "세부 정보를 모른다."



고려대학교 KOREA UNIVERSITY

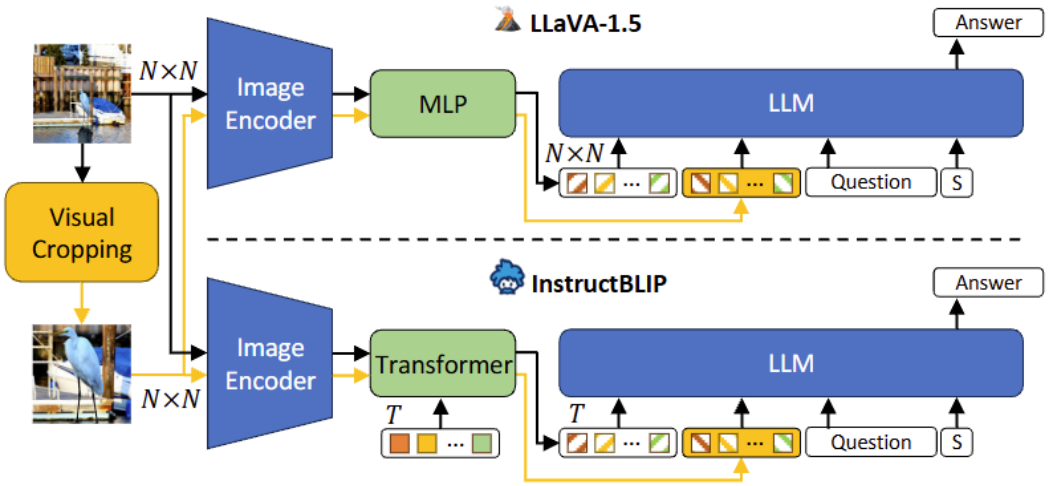
Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR) – 어떻게 작은 객체를 잘 볼 수 있을까?

- **앞선 실험에서 Cropping 효과 입증:** 원본 이미지와 Cropping한 이미지를 함께 입력하자.

Model	Method	Answer Bbox Size (S)		
		small	medium	large
BLIP-2 (FlanT5$_{XL}$)	no cropping	12.13	19.57	36.32
	human-CROP	55.76	52.02	45.73
InstructBLIP (Vicuna-7B)	no cropping	21.79	30.58	45.30
	human-CROP	69.60	61.56	53.39
LLaVA-1.5 (Vicuna-7B)	no cropping	39.38	47.74	50.65
	human-CROP	69.95	65.36	56.96
Qwen-VL (Qwen-7B)	no cropping	56.42	65.09	68.60
	human-CROP	70.35	75.49	71.05
GPT-4o	no cropping	65.76	72.81	69.17
	human-CROP	75.63	81.32	71.72



고려대학교 KOREA UNIVERSITY

Algorithm

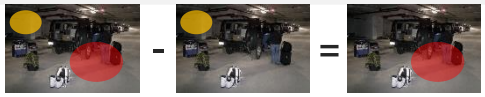
- (2) 입력 이미지 개선 기반 접근

❖ **ViCrop (2025/ICLR) – 어떻게 이미지마다 최적으로 Cropping을 어떻게 할까?**

- **Rel-att:** 일반 지시문 대비 질문에서 더 주목한 영역을 강조
- **Grad-att:** 답변에 실제로 기여한(Gradient > 0) Attention만 강조
- **pure-grad:** Attention 대신 입력 이미지의 Gradient로 중요 영역 탐색

① rel-att
Relative Attention

질문 Attn ÷ 일반 지시문 Attn



질문 | 일반 지시문 | Relative

▼

질문과 무관하게 늘 높은 Attention 제거

Forward 2회

② grad-att
Gradient-weighted Attention

Attention × Gradient (+)



Attention | Gradient(+) | 결과

▼

Attn과 Gradient를 동시에 고려

Forward + Backward 1회

③ pure-grad
Input Gradient

입력 이미지의 Gradient



입력 Gradient | Edge Mask | 결과

▼

Edge 필터로 배경 노이즈 제거
Attention 구조와 무관 → 모든 MLLM

Backward 1회

***gradient:** Grad-CAM과 같은 Sailency Map을 의미

고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR) – 실험결과 (예측 성능)

- 작은 물체 벤치마크(TextVQA, V*, POPE, DocVQA) 학습 없이 큰 폭으로 성능 향상
- 큰 물체 벤치마크(AOKVQA, GQA, VQAv2)에서도 성능 유지 → 기존 능력 손상 X
- 주로 Attn 기반 방법론(rel-att & grad-att)들이 효과적

Model		Smaller Visual Concepts				Larger Visual Concepts		
		TextVQA†	V*	POPE	DocVQA	AOKVQA	GQA	VQAv2
LLaVA-1.5	no cropping	47.80	42.41	85.27	15.97	59.01	60.48	75.57
	rel-att	55.17	**62.30**	**87.25**	19.63	**60.66**	60.97	**76.51**
	grad-att	**56.06**	57.07	87.03	**19.84**	59.94	**60.98**	76.06
	pure-grad	51.67	46.07	86.06	17.70	59.92	60.54	75.94
InstructBLIP	no cropping	33.48	35.60	84.89	9.20	60.06	49.41	76.25
	rel-att	45.44	**42.41**	86.64	9.95	61.28	49.75	**76.84**
	grad-att	**45.71**	37.70	**86.99**	**10.81**	**61.77**	**50.33**	76.08
	pure-grad	42.23	37.17	86.84	8.99	61.60	50.08	76.71

고려대학교 KOREA UNIVERSITY

Algorithm

- (2) 입력 이미지 개선 기반 접근

❖ ViCrop (2025/ICLR) – 실험결과 (예측 성능 & 효율성)

- Segmentation / Object Detection 모델보다 유의미한 영역을 우수하게 확보 → 높은 성능
- 모델 기반 대비 속도가 빠름 → 훨씬 효율적

	Model	Original	SAM	YOLO	CLIP	rel-att	grad-att	pure-grad
Accuracy (TextVQA)	LLaVA-1.5	47.80	49.42	48.84	48.55	55.17	56.06	51.67
	InstructBLIP	33.48	39.23	36.49	39.61	45.44	45.71	42.23
CPU Time	LLaVA-1.5	2.26	91.53	0.97	5.46	14.43	11.33	29.86
	InstructBLIP	0.66				4.35	3.78	7.04
GPU Time	LLaVA-1.5	0.17	3.33	0.35	1.07	1.16	0.89	2.36
	InstructBLIP	0.06				0.28	0.29	0.60

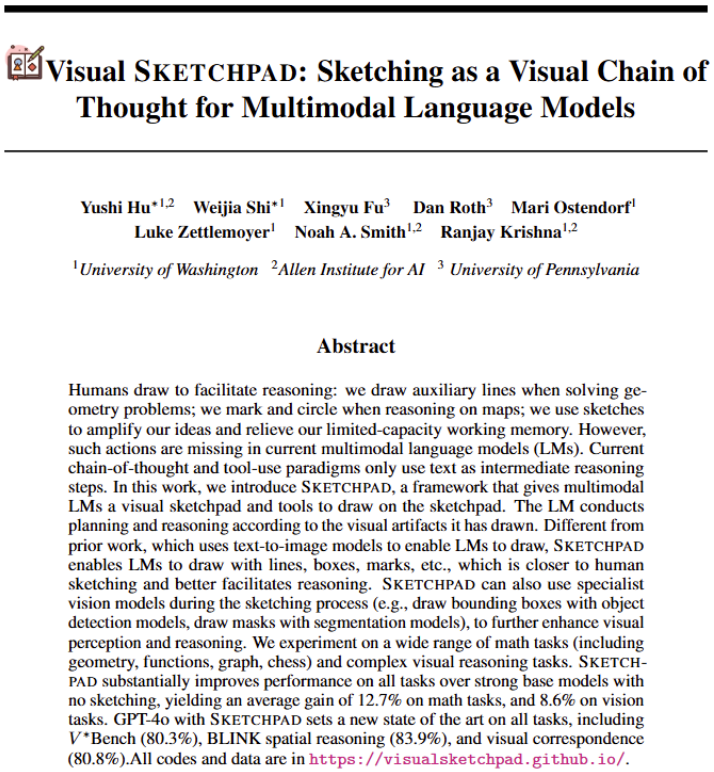
고려대학교 KOREA UNIVERSITY

Algorithm

- (3) 외부 도구를 활용한 접근

❖ Visual Sketchpad (2024/NeurIPS – UPenn)

- 인용수: 425회
- 외부 도구(tool)을 활용하여 MLLM의 Visual Reasoning 성능 개선



Visual SKETCHPAD: Sketching as a Visual Chain of Thought for Multimodal Language Models

Yushi Hu*[1,2] Weijia Shi*[1] Xingyu Fu[3] Dan Roth[3] Mari Ostendorf[1]
Luke Zettlemoyer[1] Noah A. Smith[1,2] Ranjay Krishna[1,2]

[1]*University of Washington* [2]*Allen Institute for AI* [3] *University of Pennsylvania*

Abstract

Humans draw to facilitate reasoning: we draw auxiliary lines when solving geometry problems; we mark and circle when reasoning on maps; we use sketches to amplify our ideas and relieve our limited-capacity working memory. However, such actions are missing in current multimodal language models (LMs). Current chain-of-thought and tool-use paradigms only use text as intermediate reasoning steps. In this work, we introduce SKETCHPAD, a framework that gives multimodal LMs a visual sketchpad and tools to draw on the sketchpad. The LM conducts planning and reasoning according to the visual artifacts it has drawn. Different from prior work, which uses text-to-image models to enable LMs to draw, SKETCHPAD enables LMs to draw with lines, boxes, marks, etc., which is closer to human sketching and better facilitates reasoning. SKETCHPAD can also use specialist vision models during the sketching process (e.g., draw bounding boxes with object detection models, draw masks with segmentation models), to further enhance visual perception and reasoning. We experiment on a wide range of math tasks (including geometry, functions, graph, chess) and complex visual reasoning tasks. SKETCHPAD substantially improves performance on all tasks over strong base models with no sketching, yielding an average gain of 12.7% on math tasks, and 8.6% on vision tasks. GPT-4o with SKETCHPAD sets a new state of the art on all tasks, including V^*Bench (80.3%), BLINK spatial reasoning (83.9%), and visual correspondence (80.8%).All codes and data are in https://visualsketchpad.github.io/.

[3] Hu, Y., Shi, W., Fu, X., et al. (2024). Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In NeurIPS.

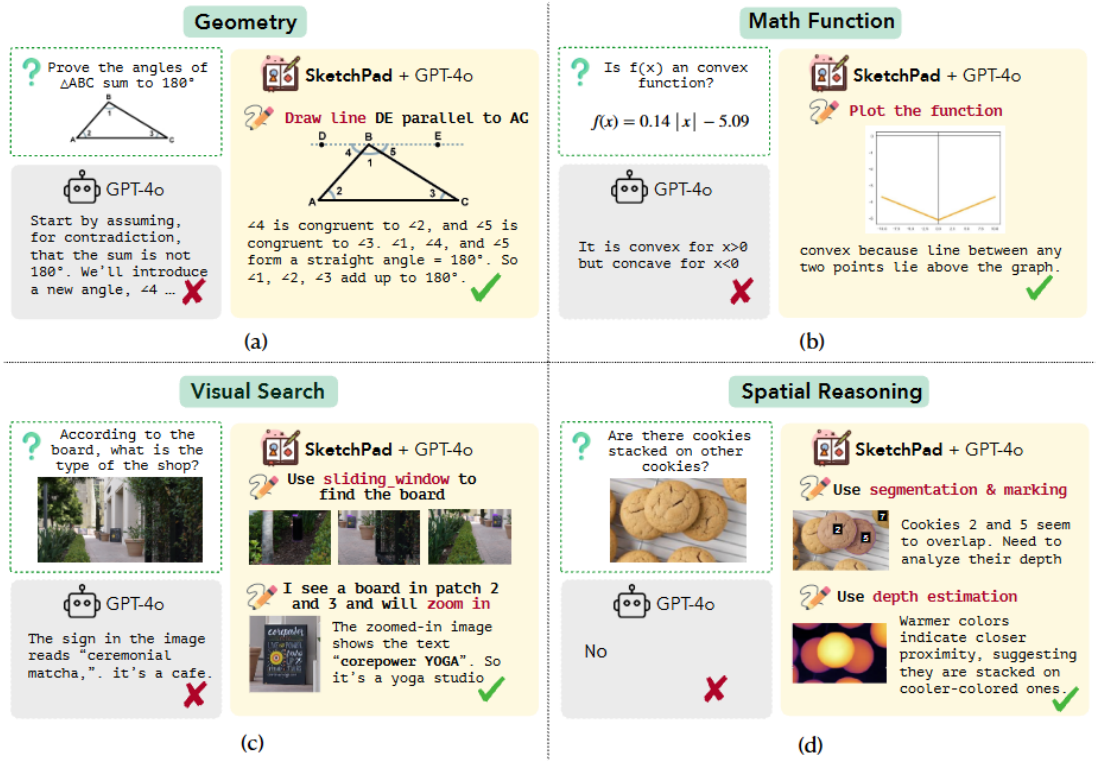
고려대학교 KOREA UNIVERSITY

Algorithm

- (3) 외부 도구를 활용한 접근

❖ **Visual Sketchpad (2024/NeurIPS) – 왜 MLLM이 Visual Reasoning에서 저조할까?**

- **인간사고:** 인간은 Visual Task를 잘 이해하기 위해 보조선을 긋고, 표시하며 시각적으로 생각
- **문제상황:** MLLM은 그런 정보 없이 주어진 입력 이미지만으로 정답을 추론하려고 함



고려대학교 KOREA UNIVERSITY

Algorithm

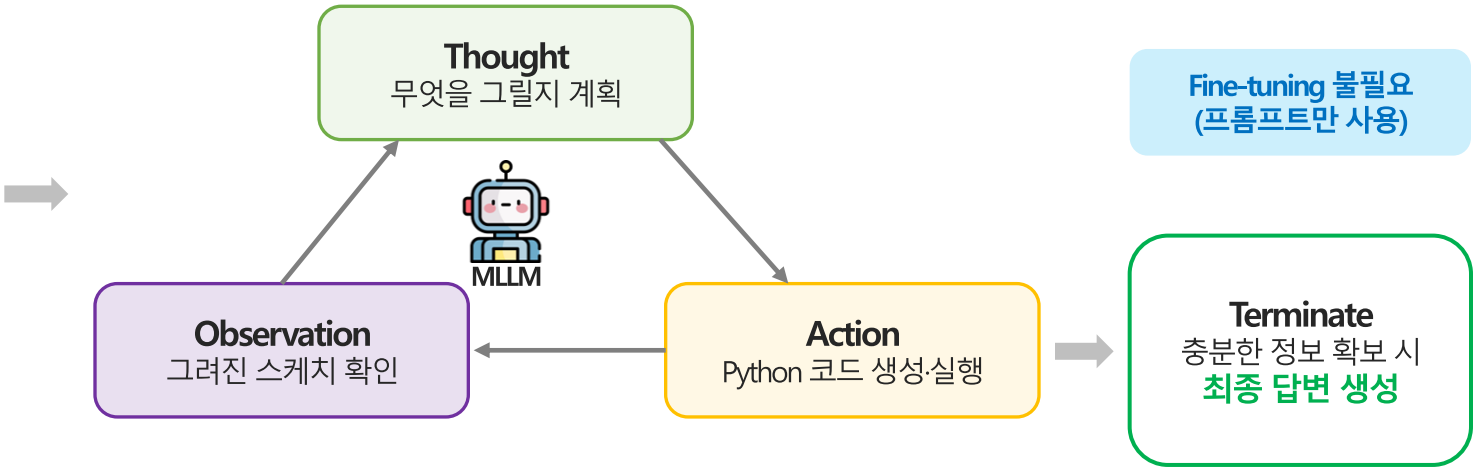
- (3) 외부 도구를 활용한 접근

❖ Visual Sketchpad (2024/NeurIPS) – 어떻게 스케치를 하며 Visual Reasoning을 할까?

- **Motivation:** 인간은 컴퓨터에서 시각화를 위해 python code 기반 도구들을 활용
- **접근방향:** 시각화에 필요한 도구들을 사전에 저장 후, 필요한 것을 골라쓰면 어떨까?



입력: 이미지 + 질문



수학	matplotlib 기하 보조선 · 함수 그래프	networkx 인접행렬 → 그래프 그리기	python-chess FEN → 체스판 그리기	display() 스케치를 Observation으로 반환
비전	detection() Grounding-DINO · BBox + 번호	sliding_window_detection() 격자 탐색으로 작은 물체 검출	segment_and_mark() SAM · Semantic-SAM 마스크 + 번호 (SoM)	
	depth() Depth Anything · 색상 깊이 맵	zoom_in_image_by_bbox() BBox 영역 Crop · 확대	overlay_images() 두 이미지 겹치기 (원본 + Depth 등)	

고려대학교 KOREA UNIVERSITY

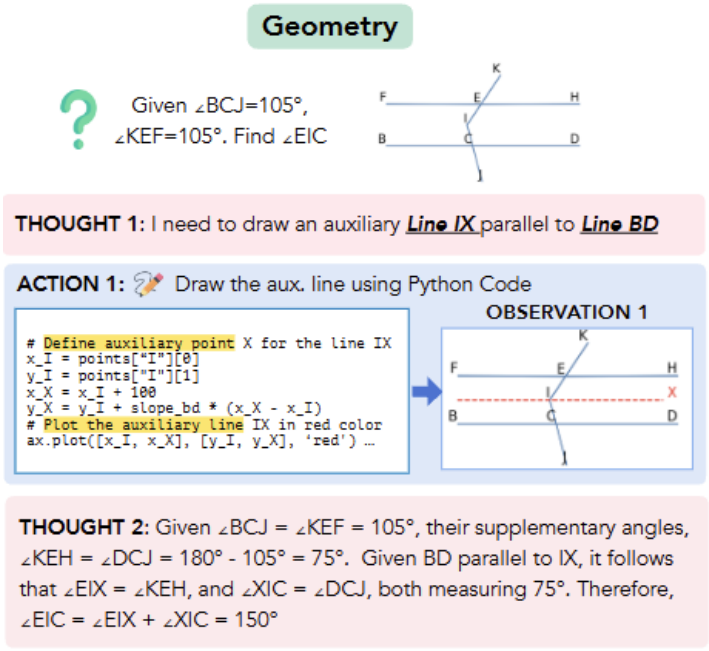
Algorithm

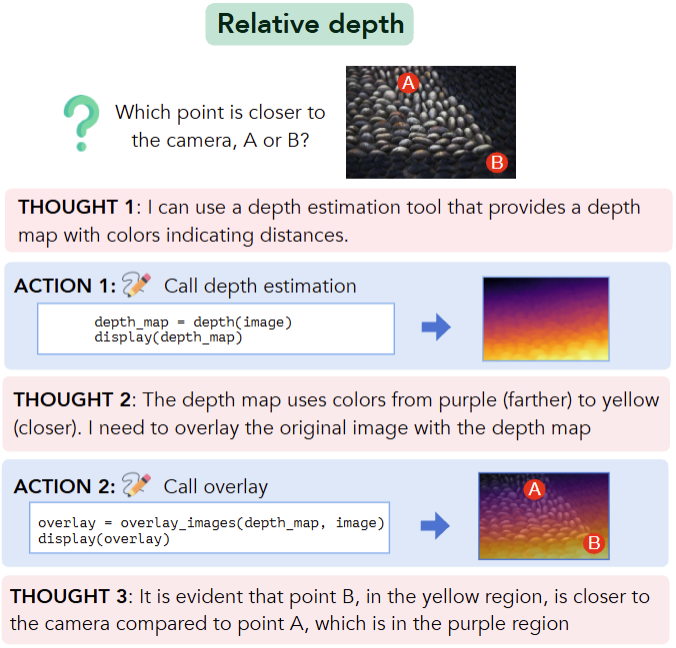
- (3) 외부 도구를 활용한 접근

❖ Visual Sketchpad (2024/NeurIPS) – 어떻게 방법론이 구성될까?

- **Thought:** 현재 맥락(질문 + 이전 단계)을 보고 다음에 무엇을 그릴지 계획
- **Action:** 계획에 맞는 Python 코드를 생성·실행하여 스케치 생성
- **Observation:** 그려진 스케치를 다시 입력받아 추론 → 충분하면 종료

  * 만약 불충분 시, 추가적으로 보조 정보 스케치 진행





고려대학교 KOREA UNIVERSITY

Algorithm

- (3) 외부 도구를 활용한 접근

❖ Visual Sketchpad (2024/NeurIPS) – 실험결과 (Main)

- 다양한 Visual Task에서 향상 → GPT-4o 기준 7개 태스크 모두 SOTA
- 단일 이미지 기반임에도, 다양한 이미지를 함께 고려하는 Jigsaw 등에서도 효과적

Model	*V**Bench	MMVP	Depth	Spatial	Jigsaw	Vis. Corr.	Sem. Corr.
Prior multimodal LLMs							
LLaVA-1.5-7B [26]	48.7	-	52.4	61.5	11.3	25.6	23.0
LLaVA-1.5-13B [26]	-	24.7	53.2	67.8	58.0	29.1	32.4
LLaVA-NeXT-34B [27]	-	-	67.7	74.8	54.7	30.8	23.7
Claude 3 OPUS [2]	-	-	47.6	58.0	32.7	36.6	25.2
Gemini-Pro [42]	48.2	40.7	40.3	74.8	57.3	42.4	26.6
GPT-4V-preview [35]	55.0	38.7	59.7	72.7	70.0	33.7	28.8
Previous state of the art	75.4 [51]	49.3 [10]	67.7 [27]	76.2 [43]	70.0 [35]	42.4 [42]	33.1 [49]
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	52.5	71.0	66.1	68.5	64.7	48.8	30.9
+ Sketchpad	71.0	73.3	68.5	80.4	68.5	52.3	42.4
	+18.5	+2.3	+2.4	+11.9	+3.8	+3.5	+11.5
GPT-4o	66.0	85.3	71.8	72.0	64.0	73.3	48.6
+ Sketchpad	**80.3**	**86.3**	**83.9**	**81.1**	**70.7**	**80.8**	**58.3**
	+14.3	+1.0	+12.1	+9.1	+6.7	+7.5	+9.7

고려대학교 KOREA UNIVERSITY

Algorithm

- (3) 외부 도구를 활용한 접근

❖ **Visual Sketchpad (2024/NeurIPS) – 실험결과 (다른 도구 기반 모델들과 비교)**

- 도구를 활용하는 방법론들 대비 Visual Sketchpad가 가장 효과적
- 다른 모델들은 Task에 따라 효과가 다름 → Visual Sketchpad는 모든 Task에서 효과적

Model	V^*	MMVP	Depth	Spatial
GPT-4 Turbo	52.5	71.0	66.1	68.5
SoM	42.0	60.7	58.9	78.3
SoM + orig.	51.3	**74.3**	66.9	79.7
Visprog	33.2	16.3	67.8	53.8
Sketchpad	**71.0**	73.3	**68.5**	**80.4**
GPT-4o	66.0	85.3	71.8	72.0
SoM	49.0	70.7	62.9	**83.2**
SoM + orig.	68.1	84.0	75.0	82.5
Visprog	32.4	17.3	46.8	37.8
Sketchpad	**80.3**	**86.3**	**83.9**	81.1

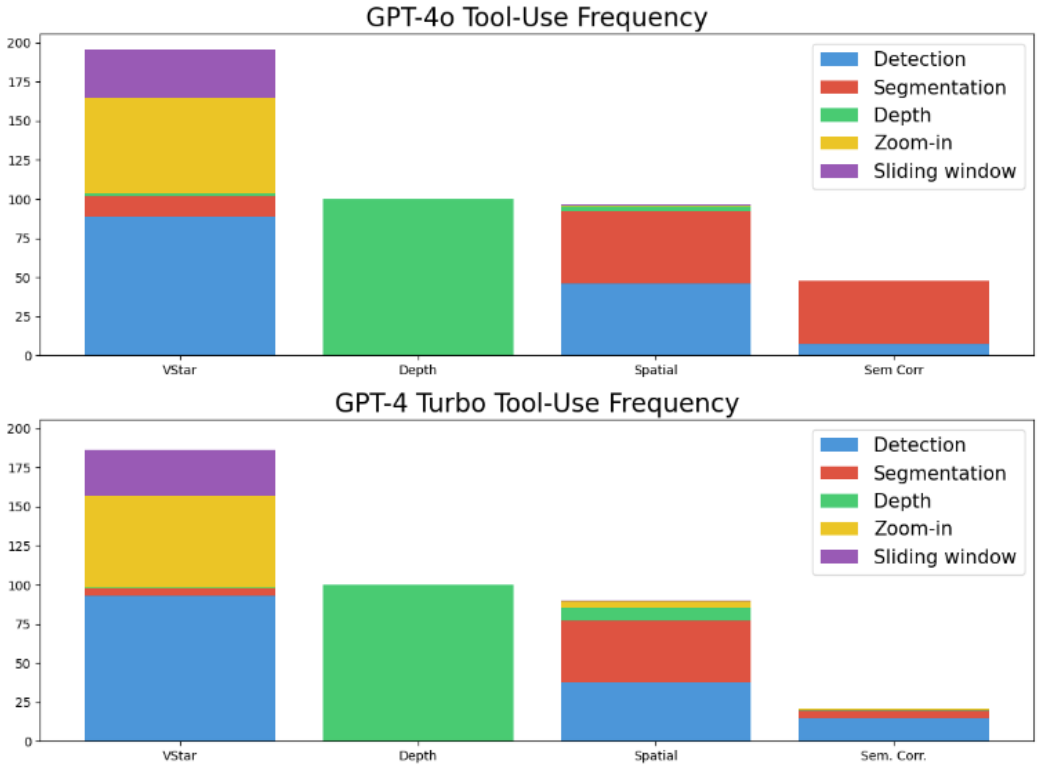
고려대학교 KOREA UNIVERSITY

Algorithm

- (3) 외부 도구를 활용한 접근

❖ Visual Sketchpad (2024/NeurIPS) – 실험결과 (Tool 사용 빈도)

- Task 별로 필요한 도구가 다름 → 이에 적합한 도구들을 효과적으로 불러옴
- Ex) 공간 추론 과제에서는 Object Detection과 Segmentation 도구를 많이 활용



고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ CoVT (2026/ECCV(Oral) – UC Berkeley)

- **인용수:** 66회
- 추가적인 학습을 통해 MLLM의 Visual Reasoning 성능 개선

Chain-of-Visual-Thought:
Teaching VLMs to See and Think Better with Continuous Visual Tokens

Yiming Qin[1,4] Bomin Wei[2] Jiaxin Ge[1] Konstantinos Kallidromitis[3]
Stephanie Fu[1] Trevor Darrell[1] XuDong Wang[1,4†]

[1]UC Berkeley [2]UCLA [3]Panasonic AI Research [4]Duke University

Project Page: https://wakalsprojectpage.github.io/covt-website/

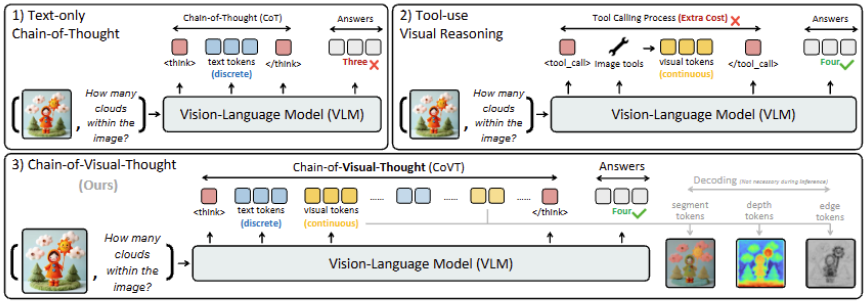


Figure 1. Traditional VLM reasoning is restricted to the discrete language space (*Text-only CoT*) or relies on costly external tool calls to obtain visual signals (*Tool-use reasoning*). **In contrast, CoVT forms a visual thought chain that enables VLMs to reason in continuous visual space.** By introducing *continuous visual tokens* that encode perceptual cues (*e.g.*, segmentation, depth, instance, and edge structure), CoVT composes *chains of textual and visual thoughts* that link semantic reasoning with perceptual grounding. These visual "thought chains" bridge language and vision, enabling fine-grained understanding, spatial precision, and geometric awareness beyond the reach of text-based reasoning. During inference, these visual tokens can be decoded into different types of visual information when interpretation is desired.

[4] Qin, Y., Wei, B., Ge, J., et al. (2026). Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens. In ECCV.

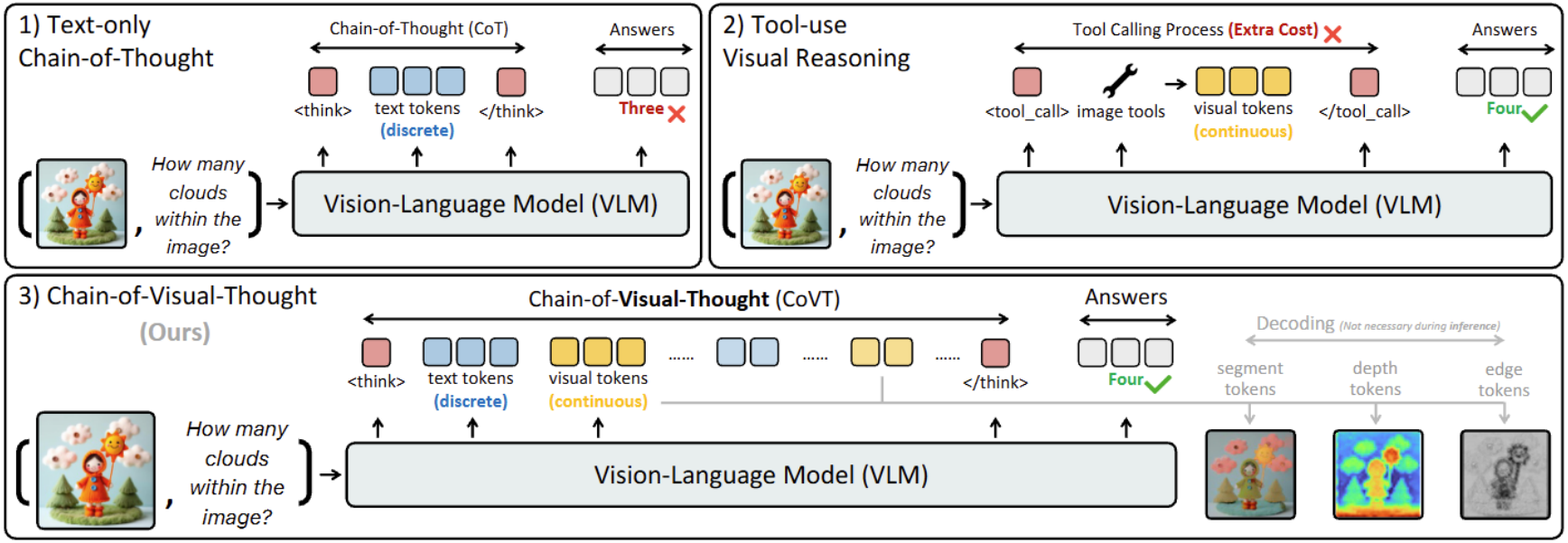
고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ CoVT (2026/ECCV) – 왜 MLLM이 Visual Reasoning에서 저조할까?

- **원인1:** Visual Reasoning을 텍스트의 이산적인 공간에서 진행
- **원인2:** 추론 시점에 외부 도구(tool)에 의존 → 많은 추론 시간과 GPU가 필요하여 비용이 큼
- **해결방향:** 연속적인 공간에서 외부 도구 활용 없이 추론하자!



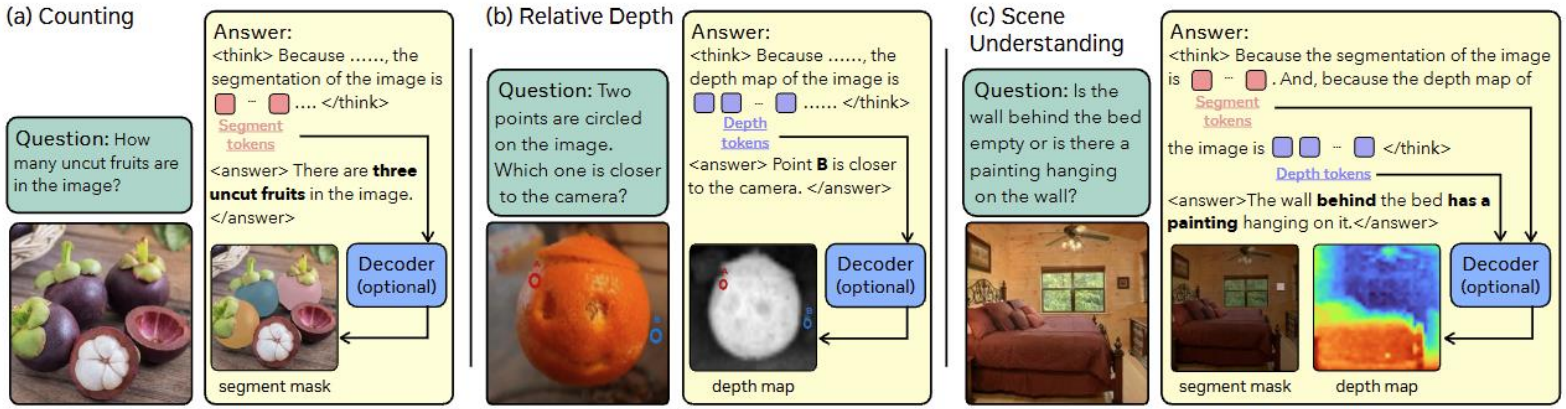
고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ **CoVT (2026/ECCV) – 어떻게 이산적인 공간이 아닌, 연속적인 공간에서 진행할까?**

- **아이디어:** 시각적인 정보를 담을 수 있는 토큰을 추가적으로 정의하자.
- 해당 토큰에 시각적 정보를 저장 → 향후 Visual Reasoning 시 해당 정보를 토대로 추론 가능



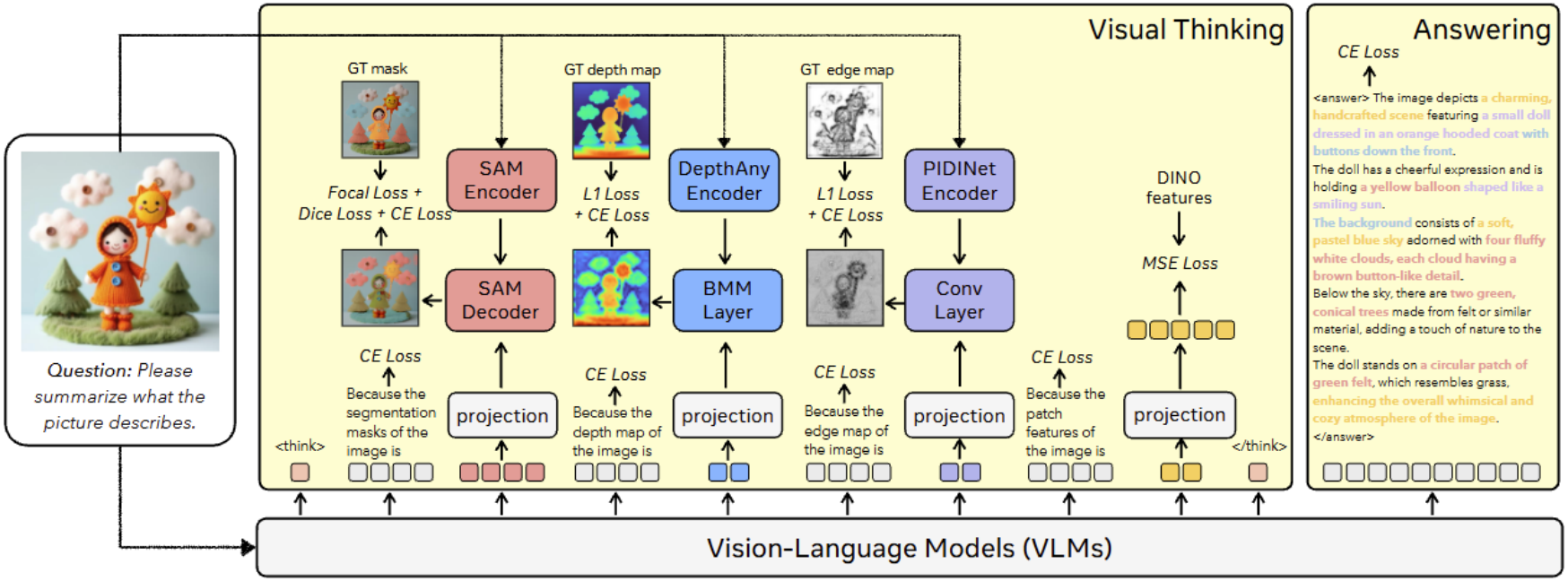
고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ **CoVT (2026/ECCV) – 어떻게 외부 도구 없이 Visual 지식을 잘 확보할 수 있을까?**

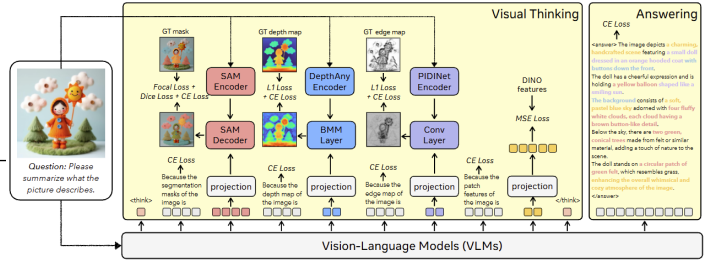
- **아이디어:** 외부 도구에 대한 지식을 추론 과정이 아닌, 학습 과정에 담자.
- **학습과정:** Segmentation (SAM) 등 4가지 Vision 전문모델을 통해 Distillation을 기반으로 학습
- **추론과정:** 외부 도구 호출 없이 단순 추론과정만으로 외부 도구 효과를 얻을 수 있음



고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근



❖ CoVT (2026/ECCV) – 시각 토큰에 어떻게 의미를 부여할까?

- **Task-level 정렬 (Seg/Depth/Edge):** Decoder를 통해 Visual Token을 복원 ←→ 실제 GT랑 비교
- **Feature-level 정렬 (DINOv2):** Visual Token을 Feature 수준에서 학습

Task-level 정렬 (SAM · DepthAnything v2 · PIDINet)

Projection Layer	SAM Decoder 토큰 1개 = Mask Prompt 1개	Mask ↔ GT (Dice+Focal)
Projection Layer	Depth Feature와 BMM 토큰 · Encoder Feature	Depth Map ↔ GT (L1)
Projection Layer	1×1 Conv Kernel 토큰을 Edge Feature에 적용	Edge Map ↔ GT (L1)

Feature-level 정렬 (DINOv2)

Projection Layer

DINOv2 Patch Feature
와 같은 형태로 맞춘 뒤 **MSE**

전체 Loss: $L_{total} = L_{CE} + \gamma (\lambda_{seg} L_{seg} + \lambda_{depth} L_{depth} + \lambda_{edge} L_{edge} + \lambda_{dino} L_{dino})$

※ VLM 모델은 LoRA로 미세조정 + Projection Layer는 모두 학습

고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ CoVT (2026/ECCV) – 실험결과 (Main & Ablation)

- 다른 방법론들 대비 성능 향상 → Seg 등 활용한 Expert 모델이 없는 Task에서도 효과적
- 3개의 Expert 모델만 활용하더라도 기존 방법론 대비 우수한 성능 확인

Visual tokens				CV-Bench				Other vision-centric benchmarks							
Seg	Depth	DINO	Edge	CVBench	Count	Depth	Dist.	BLINK	RW-QA	MMT	MMStar-P	MMVP	MME-RW	V*	HR_{4K}	HR_{8K}
Closed-source Models																
Claude-4-Sonnet				76.3	62.2	77.7	80.5	39.6	63.7	-	58.8	48.7	-	15.2	32.3	22.7
GPT-4o				79.2	65.6	86.7	81.0	63.0	69.7	-	65.2	72.0	-	42.9	50.6	46.7
Qwen2.5-VL-7B				74.5	65.0	72.8	75.5	55.7	68.6	61.7	67.1	56.0	60.0	76.4	68.6	64.9
LVR [29]†				77.2	67.0	86.2	71.2	55.9	69.7	–	67.7	–	–	81.7	69.6	65.6
Monet [49]				75.7†	**68.8**†	77.7†	73.8†	52.9†	69.3†	–	67.1†	–	–	**83.3**	71.0	68.0
CoVT (1 Visual Token)																
✓				77.9	66.0	80.8	80.5	**57.4**	71.1	62.1	68.5	58.7	62.1	79.1	71.9	69.0
	✓			78.7	65.4	83.2	78.2	56.4	71.5	**62.7**	**69.9**	58.7	62.0	79.1	71.9	69.4
		✓		71.3	64.7	72.3	66.7	55.8	71.5	62.5	67.9	57.3	61.1	77.5	71.0	68.6
CoVT (3 Visual Tokens)																
✓	✓	✓		**80.0**	66.2	86.8	**82.5**	56.0	71.6	62.1	69.2	**58.7**	**63.7**	78.0	**72.9**	69.4
Δ *(vs Baseline)*				+5.5	+1.2	+14.0	+7.0	+0.3	+3.0	+0.4	+2.1	+2.7	+3.7	+1.6	+4.3	+4.5
CoVT (4 Visual Tokens)																
✓	✓	✓	✓	79.8	66.1	**89.2**	80.5	56.2	**71.8**	61.9	68.4	56.7	63.3	78.5	72.5	**69.9**
Δ *(vs Baseline)*				+5.3	+1.1	+16.4	+5.0	+0.5	+3.2	+0.2	+1.3	+0.7	+3.3	+2.1	+3.9	+5.0

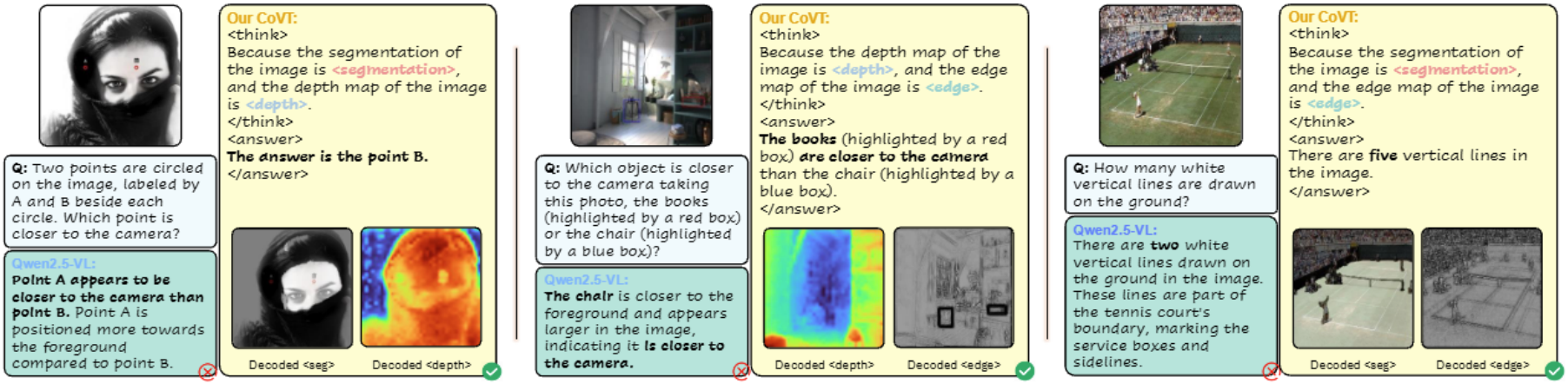
고려대학교 KOREA UNIVERSITY

Algorithm

- (4) 학습 기반 접근

❖ CoVT (2026/ECCV) – 실험결과 (추론 과정 시각화)

- 유의미한 Visual Token을 잘 활용
- 유의미한 Map을 잘 형성했음을 시각적으로도 확인

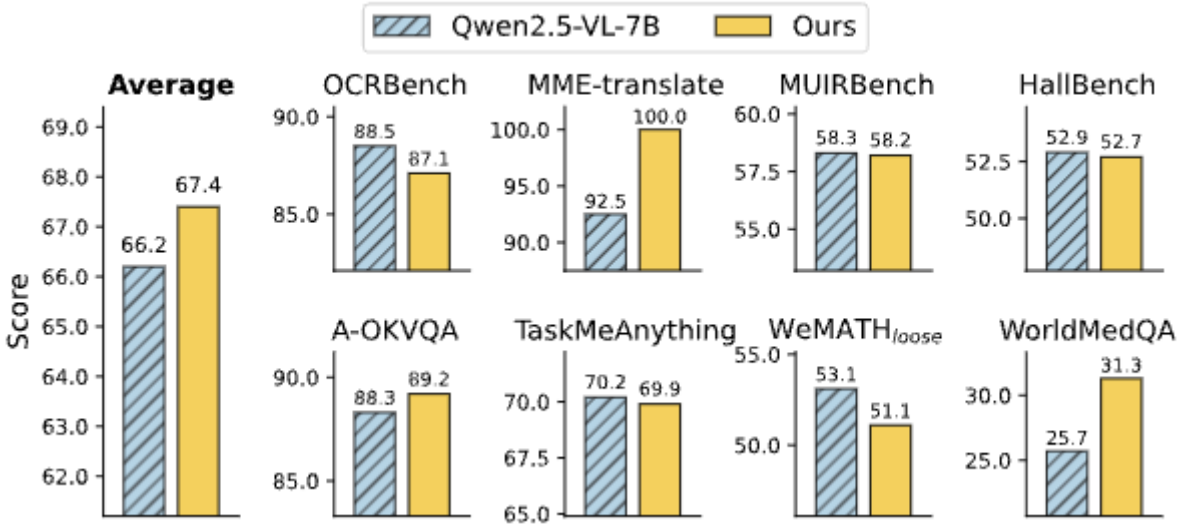


고려대학교 KOREA UNIVERSITY

Algorithm

❖ CoVT (2026/ECCV) – 실험결과 (텍스트 관련 Task)

- 텍스트 관련된 Task에서도 평균적으로 우수한 성능을 보임
- 이미지 쪽에 대해서만 추가적으로 학습했음에도, 일반 Text 관련 Task에서도 성능 거의 유지



고려대학교 KOREA UNIVERSITY

Conclusion

MLLM으로 어떻게 Visual Reasoning을 개선할 수 있을까?

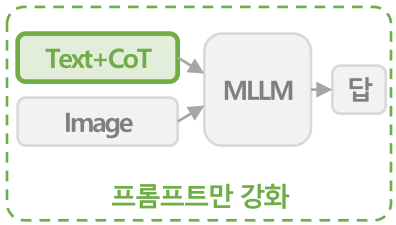
(1) 텍스트 프롬프트

CCoT

2024/CVPR

개입 지점: 텍스트 프롬프트

Scene Graph를 먼저 생성해 객체·속성·관계를 텍스트로 구조화



Training-free

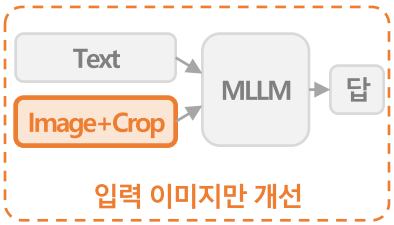
(2) 입력 이미지 개선

ViCrop

2025/ICLR

개입 지점: 입력 이미지

MLLM의 Attention으로 중요 영역을 찾아 Crop 이미지를 추가 입력



Training-free

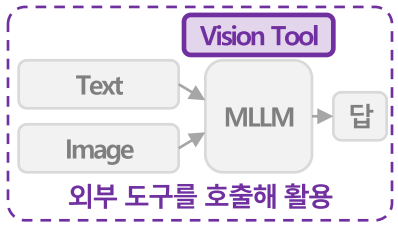
(3) 외부 도구 활용

Visual Sketchpad

2024/NeurIPS

개입 지점: 외부 비전 도구

Detection·Depth 등 도구로 스케치를 그리고, 그 결과를 보며 추론



Training-free

(4) 학습 기반

CoVT

2026/ECCV

개입 지점: 모델 내부 추론

추론 중 연속 시각 토큰을 생성해 VLM 내부에서 시각적으로 사고



학습 필요 (LoRA)

고려대학교 KOREA UNIVERSITY

Reference

1. Mitra, C., et al. (2024). Compositional Chain-of-Thought Prompting for Large Multimodal Models. In CVPR.
2. Zhang, J., et al. (2025). MLLMs Know Where to Look: Training-free Perception of Small Visual Details with Multimodal LLMs. In ICLR.
3. Hu, Y., et al. (2024). Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In NeurIPS.
4. Qin, Y., et al. (2026). Chain-of-Visual-Thought: Teaching VLMs to See and Think Better with Continuous Visual Tokens. In ECCV.

Thank you!

고려대학교
KOREA UNIVERSITY