2025 동계 세미나

Chain of thought in Multi-modal

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

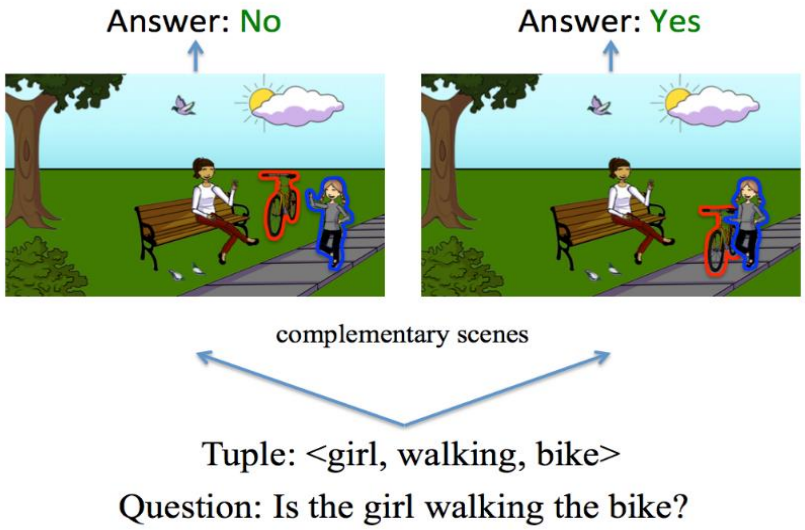
Presented By
이혜빈

Contents

- Paper review
  - Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning (NeurIPS 2025 Spotlight)
  - Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning (AAAI 2024)

Background

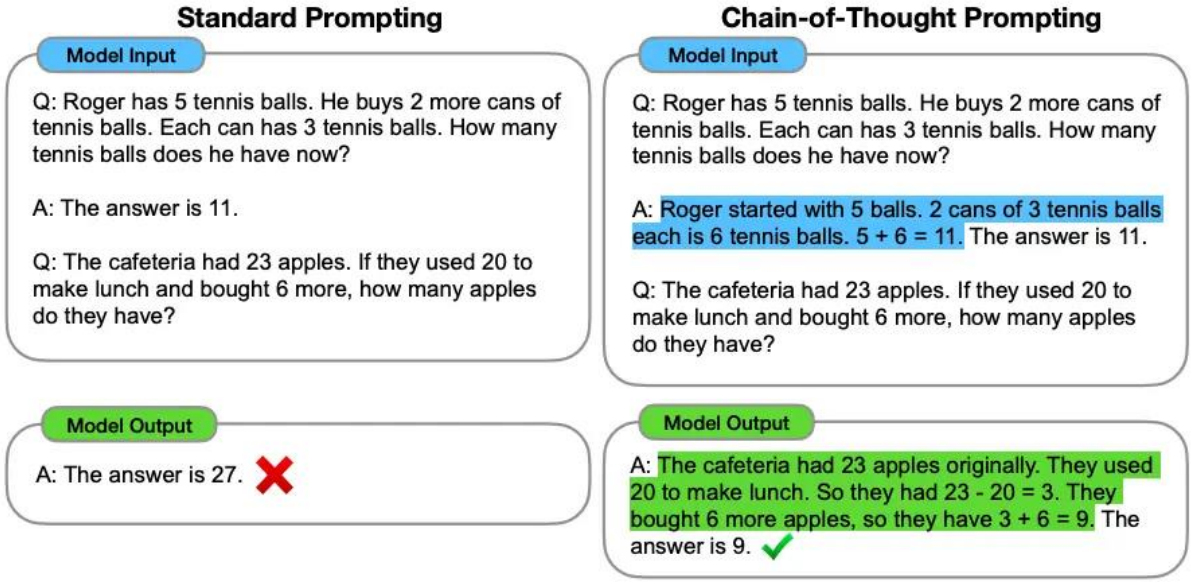
- Visual Question Answering (VQA)
  - 모델이 이미지를 이해하고 질문을 처리
  - Input: 이미지, 이미지에 대한 question
  - Output: 질문에 대한 answering



- Neural-Symbolic Visual Reasoning
  - Neural network가 이미지에 대해 이해하고 symbolic한 이해를 통해 추론함
    - Symbolic한 이해: 사람이 물체를 정의하는 방식
      - 빨간 동그란 과일을 보고 "사과" 라고 부르자고 약속하는 방식
    - 질문과 답변 과정이 XAI(설명 가능한 AI) 방식이므로 기존 black box 방식과 다름
      - 모델이 어떤 결정을 했는 지 추론할 수 있음
    - Neural network의 이해와 symbolic한 논리 구조를 결합해야 함
      - 실제 데이터셋에서 성능이 낮은 경향을 보임

Background

- Chain of Thought[1)]
  - Standard Prompting
    - (Question, Answer) prompting으로 LLM의 output을 도출
    - 수학 문제의 경우, 일반적으로 LLM이 answering을 잘 하지 못함
  - Chain-of-Thought Prompting
    - (Question, 풀이 과정이 포함된 answer) prompting으로 LLM의 output을 도출→ 성능 향상



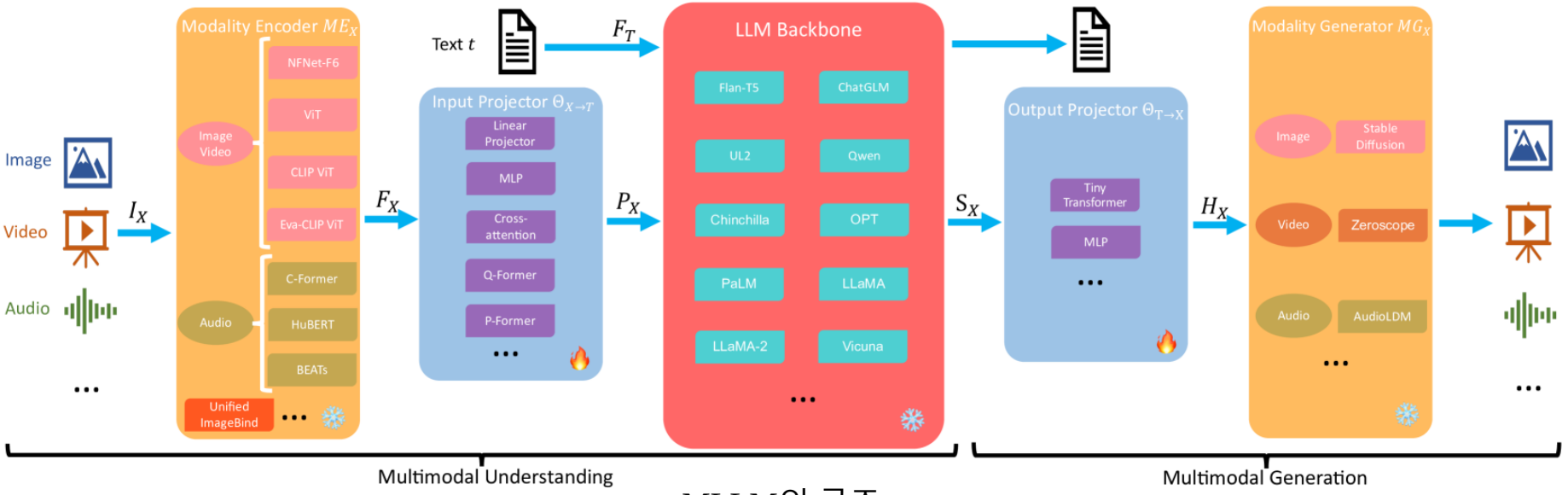
<Chain of thought 예시>

Background

- MLLM
  - Multi modal large language model
  - Text나 Image 또는 audio 등의 다양한 modal을 처리할 수 있는 모델을 의미함
    - 대표적인 예시: CLIP, GPT-4V, Flamingo, …
  - 다양한 modal에서 LLM backbone의 추론을 활용
  - LLM의 방대한 지식으로 인해 VQA 같은 다양한 task에서 뛰어난 성능 보임

Background

- MLLM 구조
  - Input image에서 visual token 추출 후, 언어적인 modality와 정렬
  - LLM 내에서 visual token, language token을 함께 처리하는 방식으로 동작
  - Modality generator에서 auto-regressive 방식을 사용
    - 출력의 이전 값을 기반으로 다음 값을 예측하는 방식
    - Black box 형태로 동작하며 시각적 입력에 기반한 명령에 응답하도록 학습되어 있음



<MLLM의 구조>

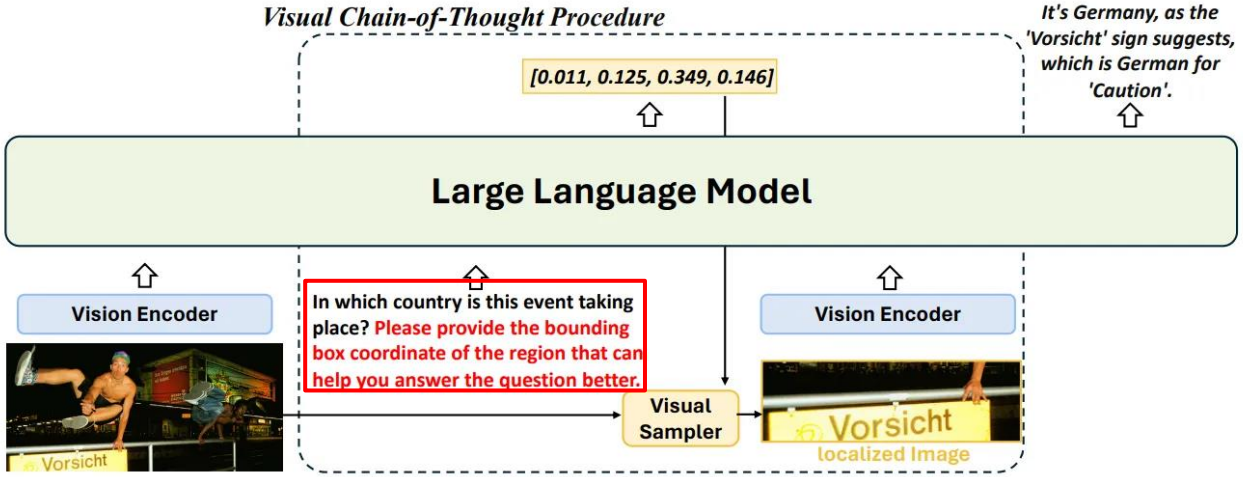
Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning (NeurIPS 2024 Spotlight)

Introduction

- Visual CoT
  - MLLM의 한계
    - MLLM이 black box의 형태이므로 해석 가능성이 떨어지는 문제가 존재함
    - Input resolution이 높거나 질문 핵심 영역의 ROI가 작으면 시각적인 입력 처리가 어려움
      - MLLM이 작은 영역에 대한 집중을 잘 하지 못함
  - 논문이 제안하는 것
    - Multi-tern 처리 파이프라인
      - 질문에 대한 답을 이해하고 확인하는 과정을 반복적으로 수행하는 방식
    - 특정 object 식별을 요구하는 VQA에서 MLLMs를 평가하기 위한 관련 benchmark
      - Input image에서 특정 object에 대한 질문을 하면 해당 object를 image에서 찾아 답변하는 방식
    - Visual CoT 데이터셋, pre-trained model

Method

- Visual CoT framework
  - 사람의 이미지 처리 방식을 모방
    - 사람: 전체 이미지를 스캔 후, 보고자 하는 영역에 집중
    - Visual CoT: 전체 이미지를 모델이 scan, 집중할 영역을 crop 후, 두 정보를 활용
  - 다음의 text를 prompt에 추가하여 LLM의 input으로 넣게 됨
    - "답변을 더 잘 제공할 수 있는 영역의 bounding box 좌표를 제공해 주세요."
    - LLM을 통해 얻은 bounding box를 통해 집중할 영역에 대해 crop하게 됨
    - 전체 이미지와 crop 이미지를 input으로 사용하여 LLM을 통해 추론



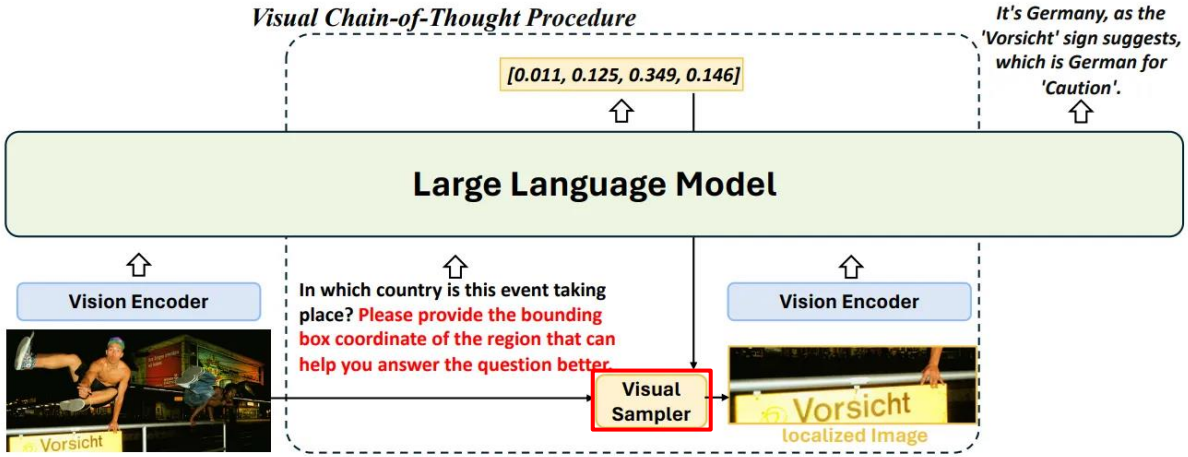
<Visual CoT framework>

Method

- Visual CoT framework
  - Visual CoT에 대한 annotation이 없으면 global 이미지만 사용하게 됨
  - Visual Sampler
    - 관심 object에 대해 정확하게 crop하는 역할 수행함
    - CLIP의 정사각형 receptive field와 정확한 object 영역 crop 위해 다음의 샘플 크기 설정

$$s = \max\{\max\{w_{half}, h_{half}\}, res_{half}\}$$

  - Vision encoder: CLIP (ViT-L/14)
  - LLM: Vicuna-7/13B
    - LLaMA 모델을 기반으로 한 대화형 LLM, 다양한 질문에 자연스러운 답변을 생성함



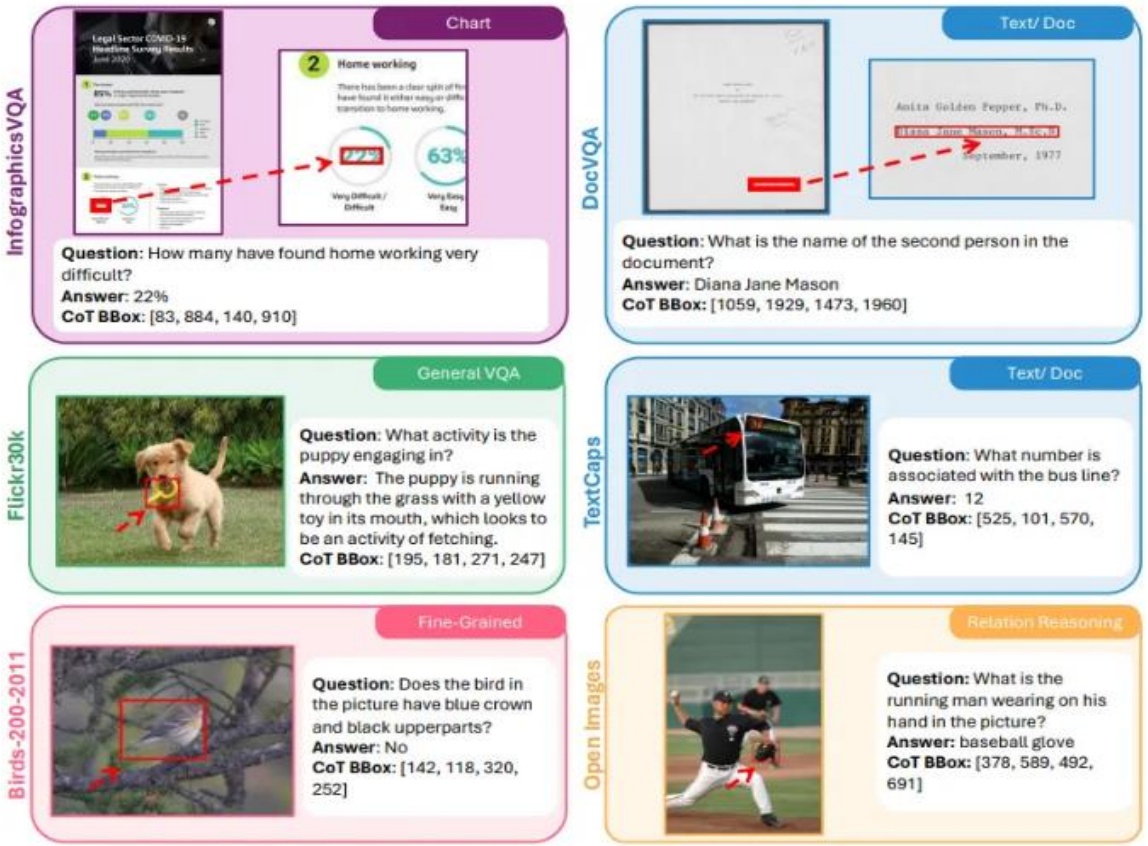
<Visual CoT framework>

Method

- Visual CoT framework
  - 학습 방법
    - 1 stage
      - 1 epoch로 model을 pre-train
      - Vision encoder와 LLM의 weight를 freeze 후, image-text caption data로 학습
        - ✓Two-layer MLP vision-language connector만을 학습
        - ✓Image-text caption dataset
          - LLaVA-1.5, Shikra
    - 2 stage
      - 해당 방법론의 Visual CoT dataset에 대해서 1 epoch fine-tuning
        - ✓Vision encoder, LLM 등 모든 가중치를 학습

Visual CoT Dataset

- 모델이 이미지 내 중요 영역을 식별할 수 있도록 bounding box annotation을 제공
- 각 데이터 샘플은 질문, 답변, bounding box 포함
  - 일부 데이터 샘플에는 상세한 추론 단계가 포함



Visual CoT Dataset

- Visual CoT dataset을 구성하는 데이터셋들의 task
  - Text/Documents: OCR, 맥락 이해 능력 향상
  - Fine-grained understanding: 시각적인 외형, 패턴에서 미세한 차이 구별
  - Chart: 시각적 데이터 해석 능력 향상
  - 일반적인 VQA: 다양한 시각적인 질문을 모델에게 해서 일반적인 사용성 향상
  - Relation reasoning: 공간적, 맥락적 인식 능력, 탐색 능력 향상

An example of detailed reasoning steps in GQA dataset

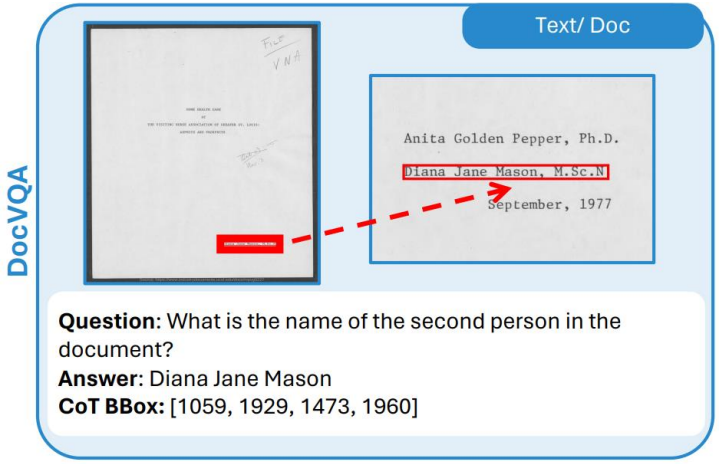
Question: What appliance is to the right of the cabinet?
###
Please think step by step and provide the bounding box coordinate of the region that can help you answer the question better.
###
Reasoning steps: 1. Identify the cabinet in the image.
2. Observe the area to the right of the identified cabinet.
3. Look for any appliance located to the right side of the cabinet.
4. Determine the name of the appliance found in this location
CoT BBox: [163, 44, 206, 67]

Answer
The appliance is a microwave.

<Reasoning 단계가 포함된 데이터 샘플>

Visual CoT Dataset

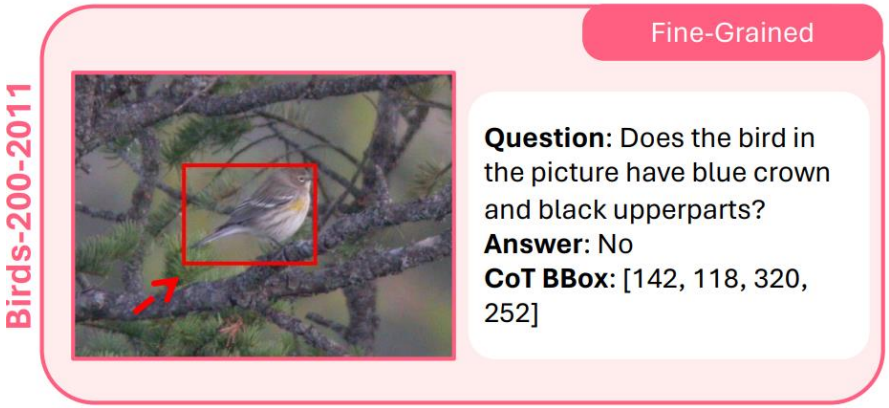
- Text/Documents
  - 텍스트 데이터셋 TextVQA, DocVQA, DUDE, TextCaps, SROIE을 사용
    - 다양한 이미지와 문서에서 텍스트 인식, 이해에 중점을 둠
    - Q-A pair 제공하는 데이터셋: TextVQA, DocVQA, DUDE, SROIE
    - 캡션, OCR 토큰만 제공하는 데이터셋: TextCaps
      - 언어 annotation tool 활용해서 적절한 Q-A를 생성
  - CoT bounding box를 생성하기 위해 PaddleOCR을 사용하여 OCR로 영역 식별
  - 답변과 일치하는 단어와 문장이 포함된 영역을 CoT bounding box로 지정
  - Bounding box로 강조된 영역이 질문과 직접적으로 관련되도록 추가적으로 필터링



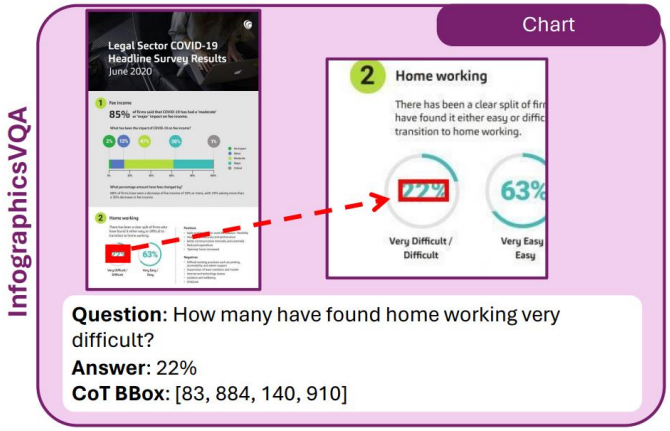
<Text/Doc 데이터 샘플>

Visual CoT Dataset

- Fine-Grained Understanding
  - Fine-grained image classification에서 사용되는 Birds-200-2011 데이터셋을 활용
    - 다양한 신체 부위 annotation, 새의 bounding box도 포함
  - MLLM에 활용하기 위해 새의 특징을 식별하도록 모델에 들어가는 질문을 작성함
- Charts
  - InfographicsVQA데이터셋을 사용
    - 고해상도의 시각화 정보를 포함, MLLM이 답변 위치를 정확하게 학습하는 데 유리함
    - 해당 방법론에서는 OCR 기술을 적용하여 답변이 포함된 영역을 식별
    - 식별된 영역을 CoT bounding box로 사용하여 모델 훈련의 정확성을 높임



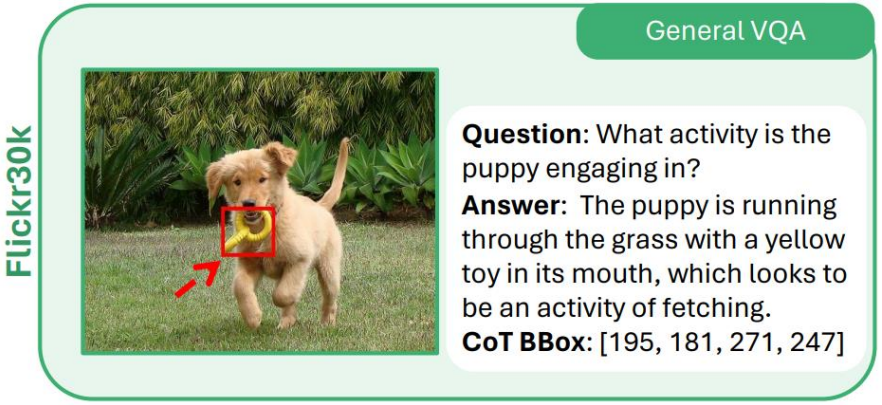
<Fine-grained 데이터 샘플>



<Charts 데이터 샘플>

Visual CoT Dataset

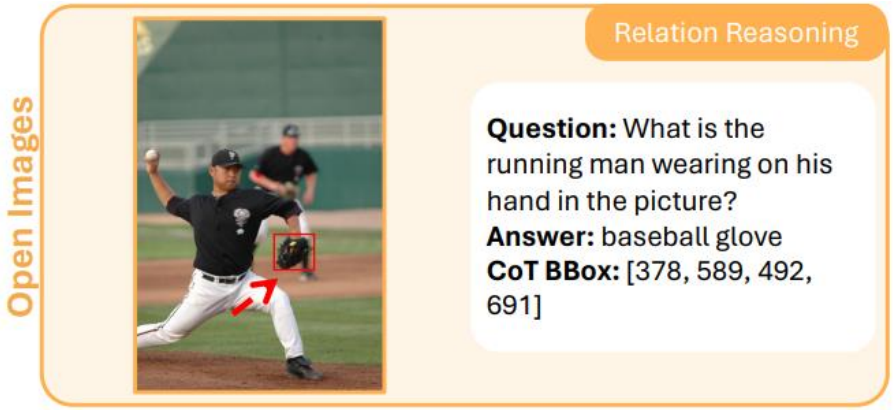
- General VQA
  - 일반적인 VQA 작업을 위해 Flickr30k와 Visual7W 데이터셋을 사용
  - Visual7W
    - Object의 bounding box와 Q-A pair를 제공
  - Flickr30k dataset
    - 이미지에 대해 5개의 caption과 caption에서 언급된 객체들에 대한 bounding box 포함
    - 해당 방법론에서는 작은 객체에 집중해야 하는 질문을 생성하기 위해 GPT-4를 활용
    - 데이터셋의 bounding box를 Visual CoT 데이터셋의 bounding box로 활용



<General VQA 데이터 샘플>

Visual CoT Dataset

- Relation Reasoning
  - Visual Spatial Reasoning (VSR), GQA, Open Images 데이터셋을 사용
    - 객체 간 공간적인 정보들이 많음
  - Visual CoT bounding box는 질문과 관련된 객체를 둘러싼 bounding box를 사용
  - 주어진 이미지와 질문을 보고, GPT-4를 사용해 더 세부적인 추론 단계를 생성함



<Relation Reasoning 데이터 샘플>

Experiments

- 평가 방식
  - 일반적인 MLLM 방법론과 동일하게 ChatGPT를 활용
    - ChatGPT에게 Question-Answer과 모델의 answer를 비교하도록 함
    - 모델의 answer가 standard answer와 가까울수록 높은 점수를 평가하도록 함
    - 해당 점수가 예측 정확도이며 0과 1 사이의 점수로 평가하도록 요청
    - 높은 점수는 더 나은 예측 정확도를 나타냄

> You are responsible for proofreading the answers, you need to give a score to the model's answer by referring to the standard answer, based on the given question. The full score is 1 point and the minimum score is 0 points. Please output the score in the form "score: <score>". The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score.
>
> Question: { }
> Standard answer: { }
> Model's answer: { }

<ChatGPT에게 평가를 요청하는 프롬프트>

Experiments

- 평가 결과
  - LLaVA-1.5와 VisCoT 모델을 테스트한 결과, Chart, Fine-grained를 제외한 모든 task에서 우수한 성능을 보였음
  - DUDE, SROIE, Visual7W는 학습 데이터에 포함되지 않았음
    - 학습 데이터셋에 포함되지 않았음에도 우수한 성능을 보임
  - Doc/Text와 Relational Reasoning task에서 압도적인 성능을 보였음

		Doc/Text					Chart
MLLM	Res.	DocVQA	TextCaps	TextVQA	DUDE	SROIE	InfographicsVQA
LLaVA-1.5-7B [39]	336^2	0.244	0.597	0.588	0.290	0.136	0.400
LLaVA-1.5-13B [39]	336^2	0.268	0.615	0.617	0.287	0.164	**0.426**
SPHINX-13B [37]	224^2	0.198	0.551	0.532	0.000	0.071	0.352
VisCoT-7B	224^2	0.355	0.610	0.719	0.279	0.341	0.356
VisCoT-7B	336^2	**0.476**	**0.675**	**0.775**	**0.386**	**0.470**	0.324

		General VQA		Relation Reasoning			Fine-grained	Average
MLLM	Res.	Flickr30k	Visual7W	GQA	Open images	VSR	Birds-200-2011	
LLaVA-1.5-7B [39]	336^2	0.581	0.575	0.534	0.412	0.572	0.530	0.454
LLaVA-1.5-13B [39]	336^2	0.620	**0.580**	0.571	0.413	0.590	**0.573**	0.478
SPHINX-13B [37]	224^2	0.607	0.558	0.584	0.467	0.613	0.505	0.419
VisCoT-7B	224^2	**0.671**	**0.580**	0.616	**0.833**	**0.682**	0.556	0.550
VisCoT-7B	336^2	0.668	0.558	**0.631**	0.822	0.614	0.559	**0.580**

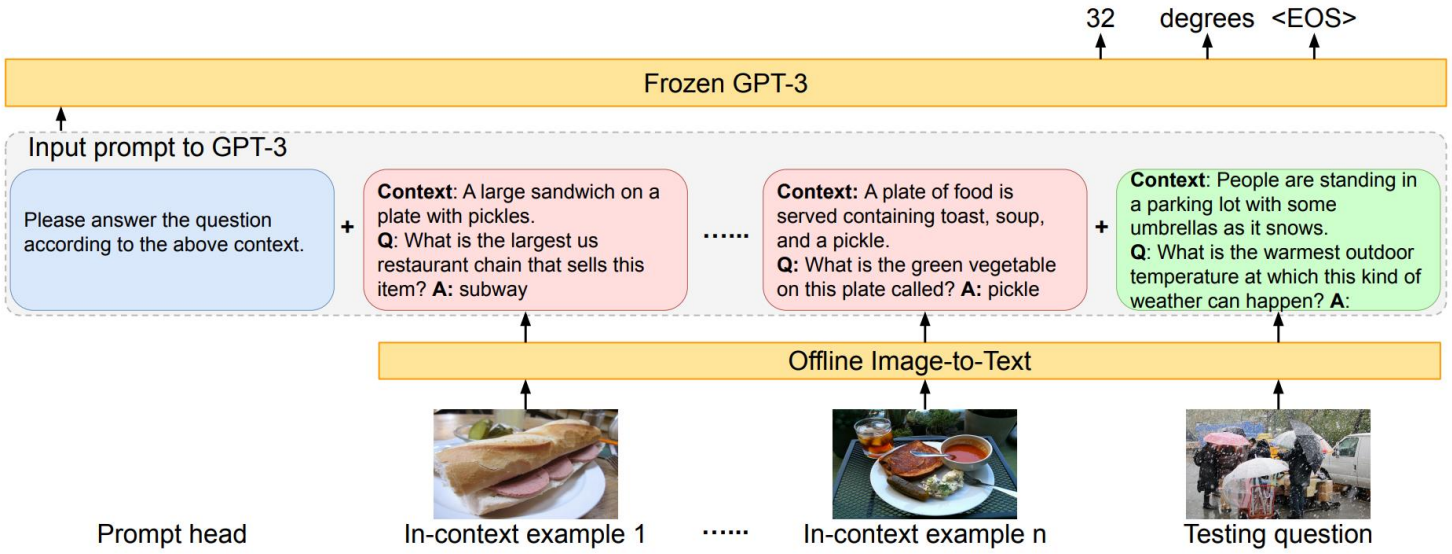
Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning (AAAI 2024)

Paper Background

- Knowledge-based visual reasoning
  - 외부적인 지식을 기반으로 시각적으로 추론하는 방식
    - Multi-modal, vision 분야에서 few-shot, fine-tuning으로의 연구
    - 이미지에 대한 caption 추출을 통해 LLM에 대한 prompt로 사용하는 방식
  - 한계
    - 대부분 LLM을 pre-trained model로 사용하여 task에 맞게 fine-tuning
      - Large model 학습 → 계산 비용 크고 시간 소모가 큼
      - 대부분 pretrained 모델은 학습한 지식 외의 질문이 들어오는 경우, 대응을 하지 못함
    - 시각, 언어 추론을 독립적으로 수행 → Modality 간의 상호 작용을 고려하지 않음

Paper Background

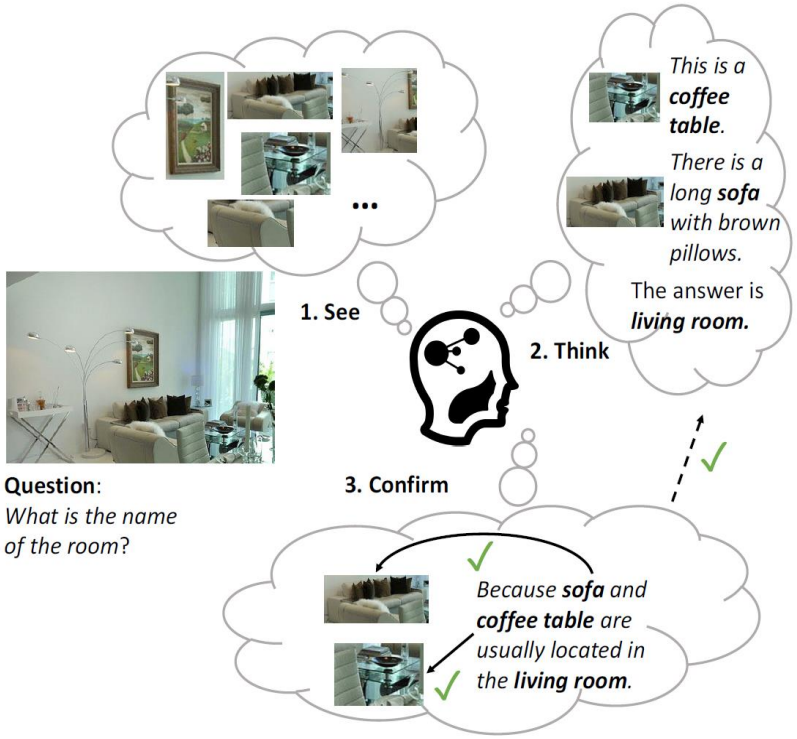
- Knowledge-based visual reasoning
  - Caption 추출 방식 (PICa[1], AAAI 2022)
    - Captioning 모델을 통해 이미지를 text 설명으로 변환
    - Text 설명과 Question을 GPT-3의 input으로 설정하여 answer를 얻는 방식을 사용하였음
    - 한계
      - 높은 성능을 달성했으나 vision 맥락과 질문이 독립적임
      - 질문-응답 과정이 블랙박스로 남아서 과정에 대한 설명이 되지 않는 문제가 존재함



<PICa의 framework구조>

Introduction

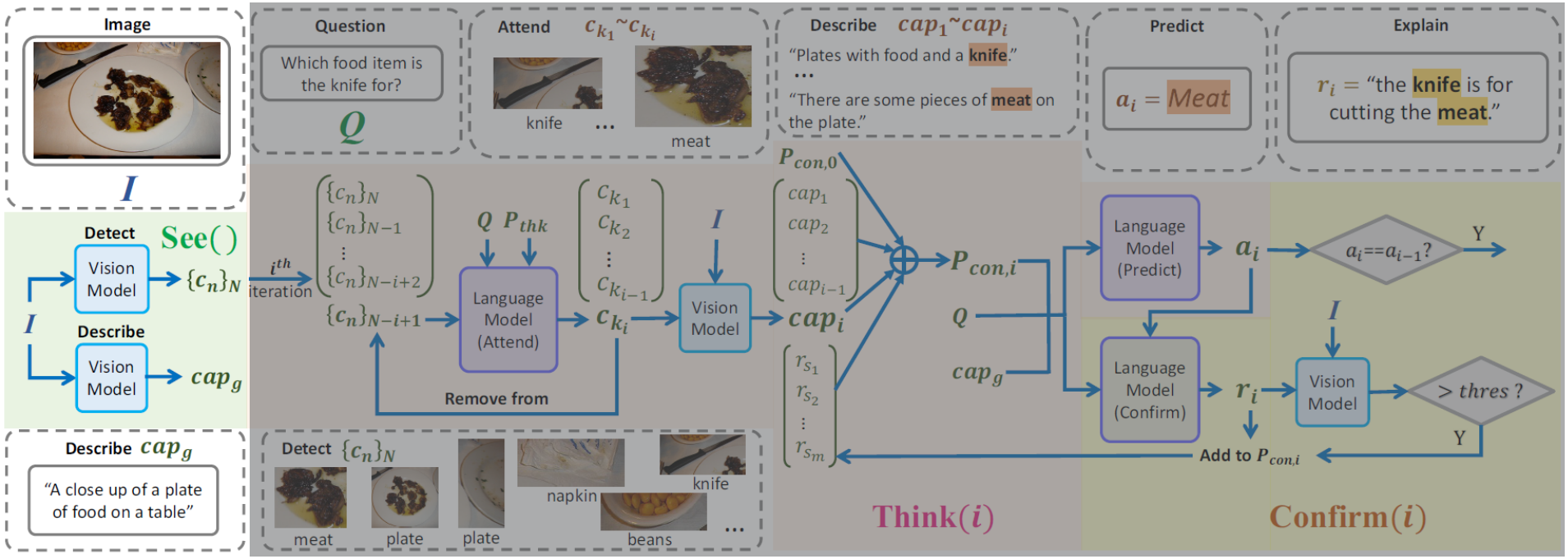
- Visual Chain-of-thought Prompting (VCTP)
  - 인간은 물체를 “보고”, “생각하고”, “확인한다” → 세 단계를 모듈화
    - See module
      - 보고자 하는 것을 특정하는 모듈
      - Image parsing
        - ✓Candidate concept 추출
        - ✓“Picture”, “Sofa”, …
    - Think module
      - 본 것에 대해서 생각하는 모듈
      - 이미지에 대한 정보를 모델이 이해해서 정답을 예측
    - Confirm module
      - 예측한 답에 대한 근거를 도출하는 모듈



  - Think 와 confirm 과정을 반복하여 답이 연속적으로 같을 때 종료됨

Method

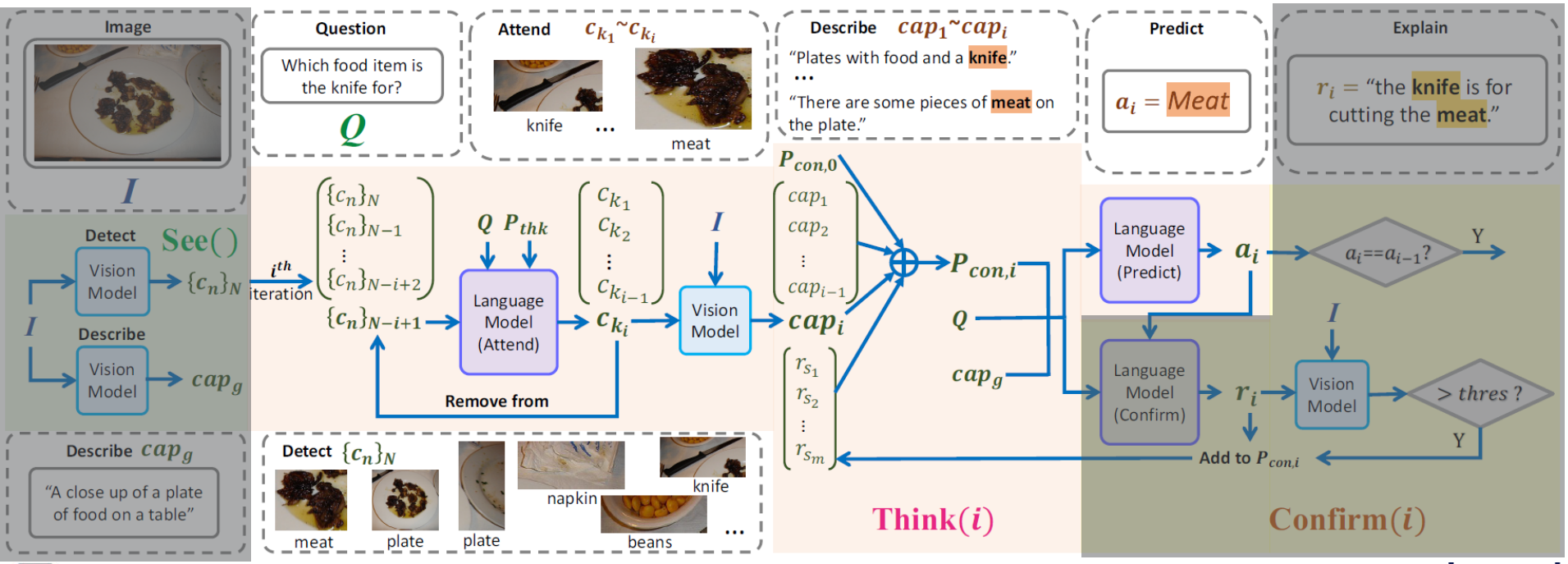
- Visual Chain-of-thought Prompting (VCTP) framework
  - See module
    - Image parser를 사용하여 image에 대한 caption(설명)과 object concept을 추출
      - Input 이미지에 대해 Faster-RCNN이 모든 candidate object를 탐지
      - Candidate object에 대한 label을 예측
    - Image captioner를 사용하여 전체 이미지에 대한 global caption을 제공



<VCTP의 framework>

Method

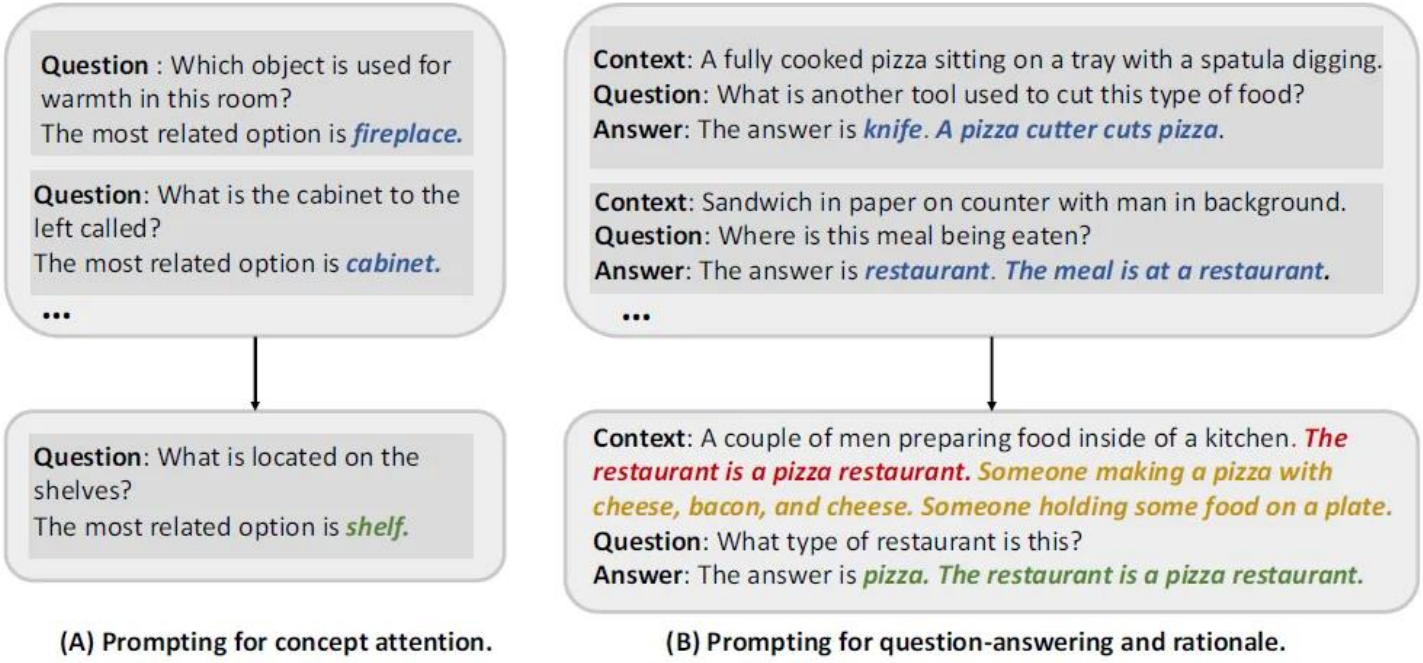
- Visual Chain-of-thought Prompting (VCTP) framework
  - Think module
    - See 모듈에서 추출한 concept(“sofa”, …)과 관련하여 LLM 을 사용해서 질문을 생성
    - Image to Caption 모델을 통해 질문에 대한 답변을 자연어 형태로 설명
    - 주의를 기울인 시각적인 context 바탕으로 LLM이 질문에 대한 답을 예측함



<VCTP의 framework>

Method

- Visual Chain-of-thought Prompting (VCTP) framework
  - Think module
    - Attend-Describe-Predict 접근 방법을 사용해서 많은 정보를 LLM에게 전달
      - Attend: 주요 concept("sofa")에 대해 LLM이 주의를 기울이도록 question을 prompting
      - Describe: Object crop image → Image to caption model → regional description 생성
      - Predict: regional description이 LLM에 context로 프롬프트에 추가되어 답변을 예측

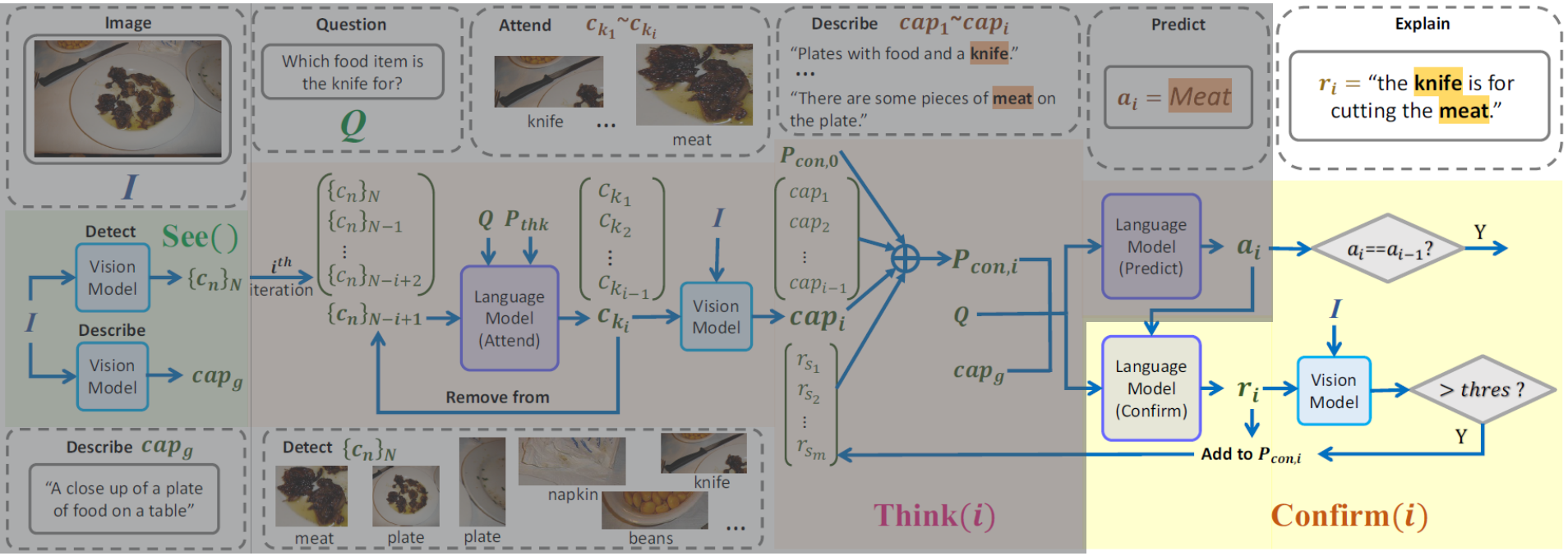


(A) Prompting for concept attention.

(B) Prompting for question-answering and rationale.

Method

- Visual Chain-of-thought Prompting (VCTP) framework
  - Confirm module
    - Prediction에 대한 근거를 생성하고 정확성 검증을 목표로 함
    - Think 모듈에서 생성된 few-shot context, interactive prompt 가 LLM의 input으로 주어짐
      - LLM은 답변 예측 후에 근거를 생성하게 됨
      - Few-shot context: regional caption, interactive prompt: question and answering prompt



<VCTP의 framework>

Method

- Visual Chain-of-thought Prompting (VCTP) framework
  - Confirm module
    - 기존 MLLM의 문제
      - 생성 process가 black box 방식이라 예측된 답변과 근거의 정확성에 대한 검증 어려움
    - 근거(rationale) 에 대한 정의를 새롭게 함으로 해결하고자 하였음
      - Rationale는 답변과 일관성이 있어야 함
        - ✓생성된 rationale을 다음 반복에서 LLM의 프롬프트에 입력
        - ✓연속된 두 답변이 같아질 때까지 반복
      - Rationale은 시각적 입력과 일관성이 있어야 함
        - ✓ Rationale이 시각적인 context와 일치하도록 하기 위해 CLIP 유사도를 사용
          - Text rationale과 image간의 matching 을 통해 검증
        - ✓유사도가 높은 rationale이 다음 반복의 prompt에 추가됨

Experiments

- Visual Chain-of-thought Prompting (VCTP) framework
  - Implementation details
    - See module에서 이미지 concept 추출 위해 Faster R-CNN을 사용
    - Think module에서 BLIP 모델을 통해 object에 대한 regional captioning 수행
    - LLM을 in-context 예제 선택과 multi-query ensemble로 prompting함
      - Input CLIP feature와 유사도가 높은 것들을 dataset에서 선택한 후 K개의 query 추출
        - ✓유사한 예시를 참조하여 모델의 성능 개선을 시도
      - K개의 query를 통해 얻은 결과를 ensemble
        - ✓Log probability가 가장 높은 결과를 선택
  - Dataset
    - OK-VQA
      - 14,055개의 이미지-질문 쌍을 포함하는 Knowledge-based VQA 데이터셋
        - ✓VQA에서 외부 지식이나 상식을 필요로 하는 유형의 데이터셋
    - A-OKVQA
      - 지식 관련 질문 뿐만 아니라 근거도 제공하여 단계별 추론을 테스트하는 데 더 적합

Experiments

- Visual Chain-of-thought Prompting (VCTP) framework
  - Metric
    - BLEU Score

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

      - Image caption 평가나 n-gram에서 주로 사용되는 평가 방식
      - BP: 생성된 문장이 짧은 경우 부여되는 penalty
      - p_n: 생성된 문장의 n-gram이 참조 문장에서 얼마나 많이 등장하는지 계산
      - w_n: 각 n-gram p_n에 대한 가중치
    - CLIP Sentence Similarity

$$\text{Similarity}(S_1, S_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}$$

      - 문장 간의 유사도를 측정하기 위해 CLIP의 텍스트 인코더를 활용
      - 문장을 벡터로 변환한 후, 코사인 유사도(Cosine Similarity)를 계산

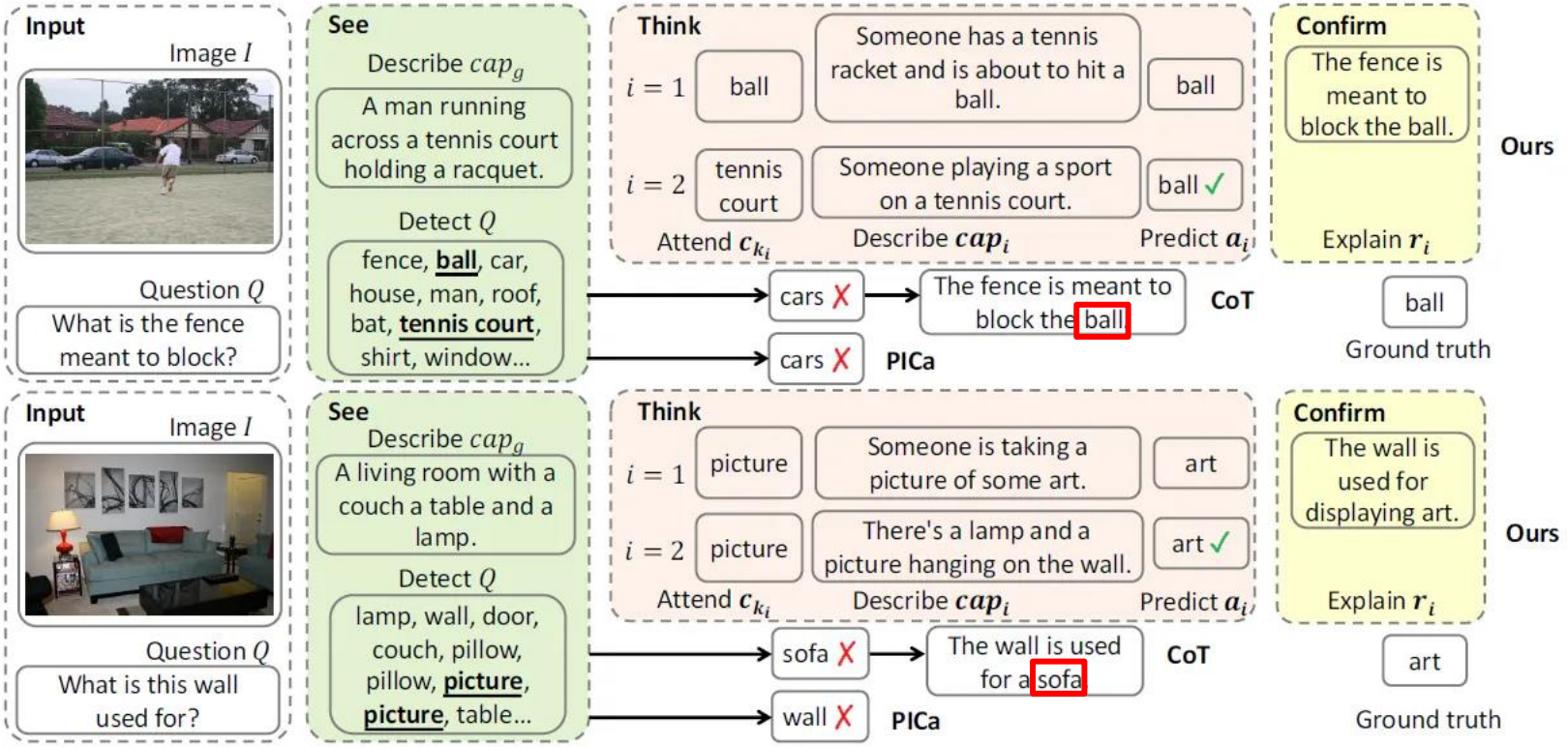
Experiments

- Visual Chain-of-thought Prompting (VCTP) framework
  - Baseline
    - PICa[1)]
      - 이미지 캡션과 객체 태그만을 사용하여 LLM(GPT3)을 프롬프트, best few-shot model
    - CoT[2)]
      - 먼저 LLM에게 근거(rationale)을 생성하도록 요청한 후, 답을 예측하도록 함
  - Quantitative Results.
    - Fully supervised model보다 성능 우수함
      - 대부분 train set에 overfitting됨
      - Validation 성능 > Test 성능
    - GPV-2에 비해 VCTP가 우수한 성능 보임

Methods	A-OKVQA Val	A-OKVQA Test	OK-VQA Test
MAVEx (Wu et al. 2022)	-	-	41.37
UnifER (Guo et al. 2022)	-	-	42.13
Pythia (Yu Jiang* et al. 2018)	25.2	21.9	-
ViLBERT (Lu et al. 2019)	30.6	25.9	-
LXMERT (Tan and Bansal 2019)	30.7	25.9	-
KRISP (Marino et al. 2021)	33.7	27.1	38.4
PICa⋆-GPT-3	-	-	48.0
GPV-2 (Kamath et al. 2022)	**48.6**	**40.7**	-
KAT-GPT-3	-	-	**54.4**
BLIP2	38.2	37.2	45.9
CoT⋆ (Wei et al. 2022)	41.5	43.7	38.1†
PICa⋆ (Yang et al. 2022)	42.4	43.8	42.9
Ours⋆	46.4	46.0	44.6‡
Ours-Llama-2⋆	50.5	**54.4**	54.9
Ours-BLIP2-Codex⋆	**53.2**	53.8	**56.2**

Experiments

- Visual Chain-of-thought Prompting (VCTP) framework
  - Quantitative Results.
    - PICa[1]와 CoT[2]에 비해, VCTP가 주요 물체에 적응적으로 답변을 생성
      - PICa[1], CoT[2]의 경우, 이미지를 통해 얻은 global caption만 사용 → 적응적이지 못함



<VCTP와 CoT, PICa의 정성적 비교 예시>

Thank you ☺