Multimodal Reasoning with Multimodal Knowledge Graph *[ACL 24]*

Junlin Lee, Yequan Wang, Jing Li, Min Zhang
Harbin Institute of Technology, Shenzhen, China
Beijing Academy of Artificial Intelligence, Beijing, China

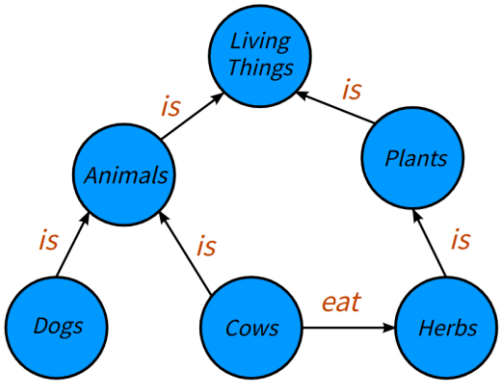
Sunghyun Ahn
skd@yonsei.ac.kr

<2025/05/27>

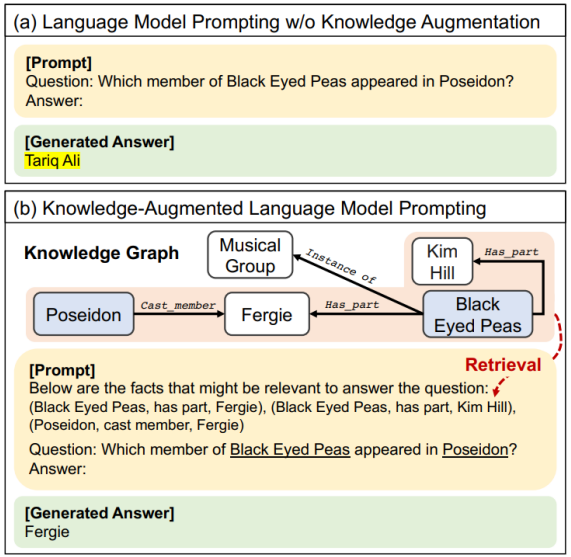
1 Introduction

Knowledge Graph

- **Knowledge Base (KB):** 두 대상의 관계를 표현한 데이터의 집합, (entity-relation-entity)으로 구성된 삼중항의 정형 데이터
- 수집한 데이터를 KB 형태로 추출한 뒤, 이를 Knowledge Graph (KG)로 변환하면 관계 구조가 시각화되고 활용성이 향상됨
- KG는 탐색, 추론 등 다양한 분야에서 활용되며, LLM과 결합하면 사전에 학습되지 않은 질문에도 정확한 답변 생성이 가능함




```
("Dogs", "is", "Animals")
("Cows", "is", "Animals")
("Animals", "is", "Living Things")
("Plants", "is", "Living Things")
("Herbs", "is", "Plants")
("Cows", "eat", "Herbs")
```



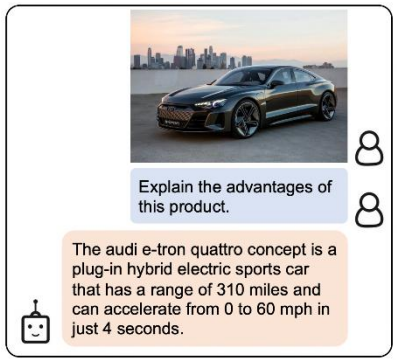
J.Baek et al. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering."ACL. 2023.

1 Introduction

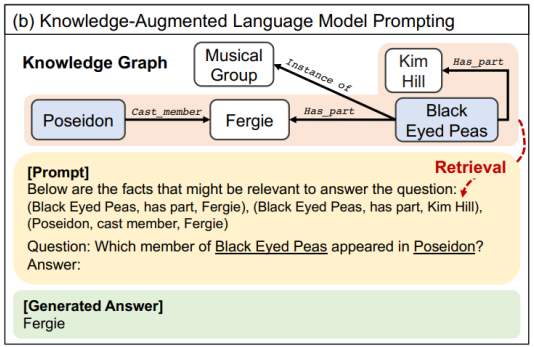
Multimodal Reasoning

- LLM이 다양한 형태의 정보를 통합적으로 이해하고 추론하는 능력
- QA를 수행하기 위해 이미지나 KG에서 중요한 정보를 추출하고, 해당 정보를 함께 활용해 정답을 찾는 과정
- 최근에는 이미지와 KG를 동시에 활용한 복합적인 질의응답(multimodal QA) 작업이 많이 연구되고 있음

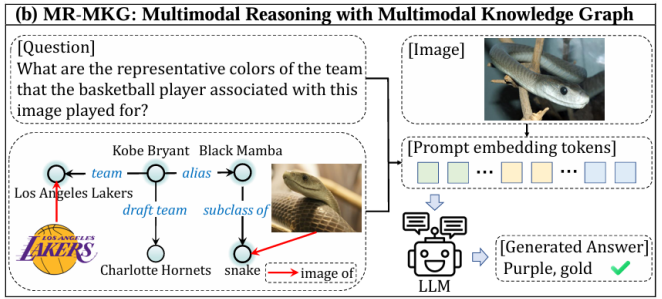
→ 제공된 이미지와 LLM의 사전지식만으로 해결하기 힘든 질문을 답변할 수 있음



Question (Text) + Image



Question (Text) + Knowledge graph



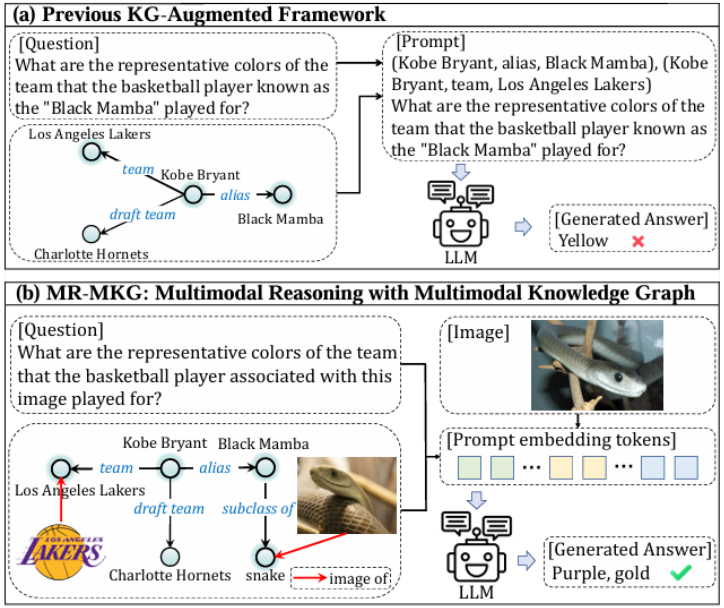
Question (Text) + Image + KG

1 Introduction

Limitation

- 단일 모달리티(text)로 구성된 KG는 시각적 정보가 필요한 QA를 수행하기 어려움
- 요구된 시각 정보가 LLM의 사전 지식에 없거나 제공된 이미지에 포함되지 않으면 hallucination이 발생할 수 있음

→ 텍스트와 이미지를 함께 활용한 Multimodal KG (MMKG)를 사용하면 다양한 형태의 질문을 더 정확하게 해결할 수 있음



1 Introduction

Multimodal Reasoning with Multimodal Knowledge Graph (MR-MKG)

- MMKG를 활용하여 LLM의 멀티모달 추론 능력을 확장하는 방법을 처음으로 제안함
- 멀티모달 정보를 LLM에 통합하기 위한 모델(MR-MKG)과 관련 데이터셋(MMKG-grounded)을 제안함
- 제안한 모델은 멀티모달 추론 과제에서 기존 기법 대비 우수한 성능을 달성함

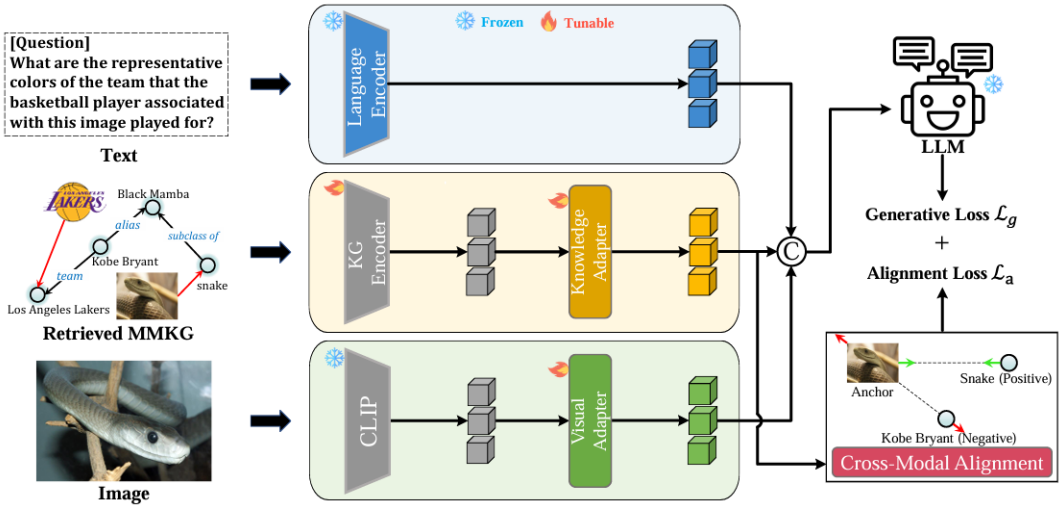


Figure 2: The overview of our MR-MKG approach. Text, multimodal knowledge graph, and image are independently embedded and then concatenated to form prompt embedding tokens. A cross-modal alignment module is designed to enhance the image-text alignment through a matching task within MMKGs.

2 Method

Overall Architecture

- 각 모달리티의 임베딩을 추출하기 위해 세 개의 인코더 구조(Language, KG, Image)를 사용함
- 세 임베딩은 결합되어 LLM의 입력(prompt)으로 사용되며, 이를 통해 LLM은 멀티모달 정보를 활용 가능함
- MMKG에 위치한 이미지-텍스트 임베딩 간 매칭을 학습하여 성능을 향상시킴 (Cross-Modal Alignment)

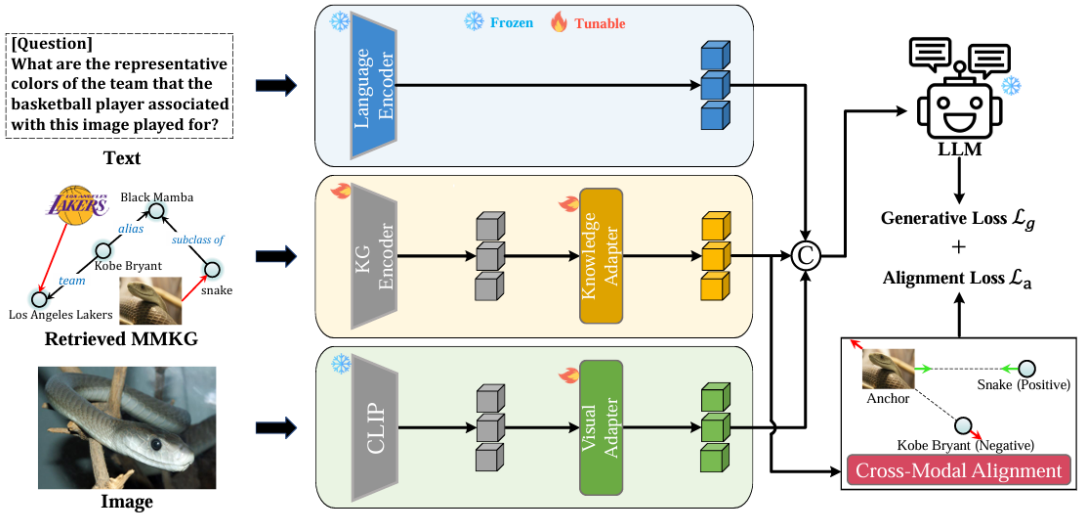
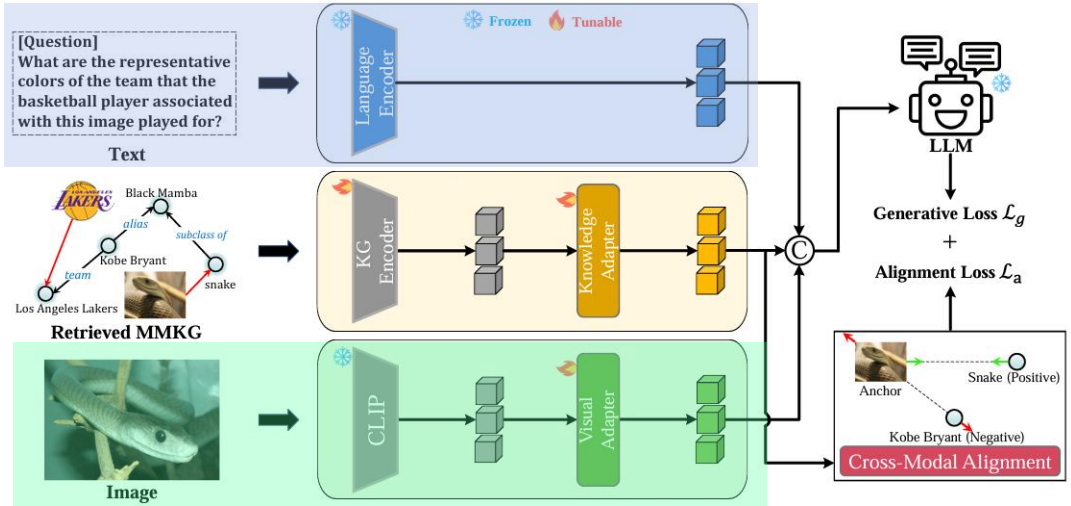


Figure 2: The overview of our MR-MKG approach. Text, multimodal knowledge graph, and image are independently embedded and then concatenated to form prompt embedding tokens. A cross-modal alignment module is designed to enhance the image-text alignment through a matching task within MMKGs.

2 Method

Language and Image Embedding

- 언어 임베딩은 사전학습된 LLM의 embedding layer를 인코더로 활용함
- 이미지 임베딩은 사전학습된 CLIP의 이미지 인코더를 사용함
- 이미지 임베딩은 Adapter를 통해 LLM이 이해할 수 있는 표현으로 변환되며, 이 과정에서 어텐션 연산을 통해 Question과 관련된 영역을 강조함



$$H_I = W_I \cdot X_I + b_I \quad (1)$$

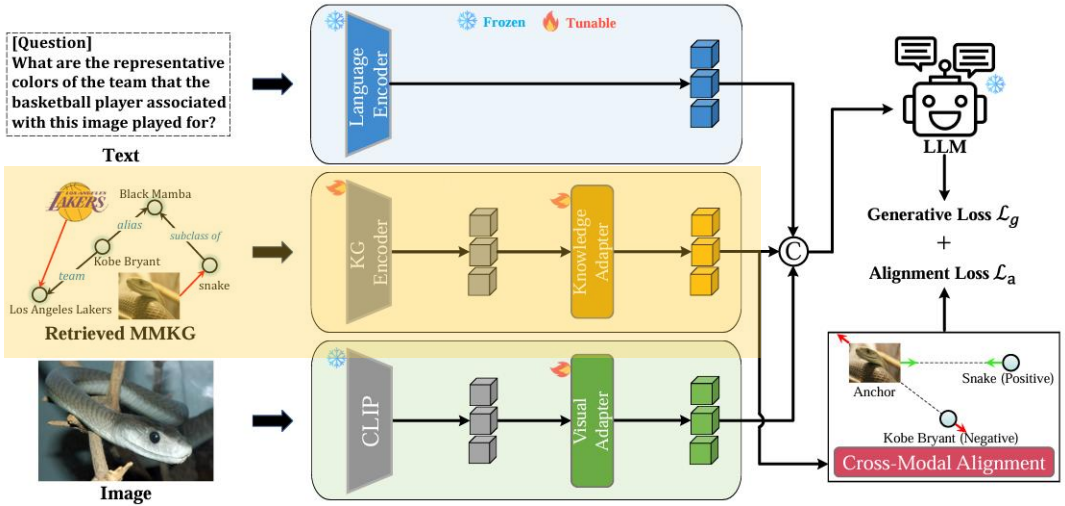
$$H_I' = \text{Softmax}(\frac{H_T H_I^\top}{\sqrt{d_k}}) H_I \quad (2)$$

where d_k represents the dimension of H_T, and W_I represents the trainable visual adapter matrix.

2 Method

KG Embedding

- KG 임베딩은 관계 기반 그래프 어텐션 모델인 RGAT를 활용함 (노드 및 엣지 임베딩은 CLIP을 통해 초기화됨)
- RGAT는 관계 유형과 이웃 노드의 특성을 함께 고려하여, 중요한 이웃 정보를 선택적으로 집계함
- 집계된 KG 임베딩은 선형 변환 및 어텐션을 통해, Question과 관련된 핵심 노드 정보를 강조함



$$\mathbf{h}_{ir}^{(l-1)} = \sum_{j \in \mathcal{N}^r(i)} \alpha_{ijr}^{(l-1)} W_r^{(l-1)} \mathbf{h}_j^{(l-1)} \quad (4)$$

$$\mathbf{h}_i^{(l)} = \sum_{r=1}^{R} \mathbf{h}_{ir}^{(l-1)} \quad (5)$$

where $W_r^{(l-1)}$ denotes a learnable weight matrix for each relation r. In addition, We apply multi-head attention to the aggregation module in (4) and concatenate its outputs. After this propagation module in (5), we use layer normalization with learnable affine transform parameters.

$$H_K = W_K \cdot X_K + b_K \quad (4)$$

$$H'_K = \text{Softmax}(\frac{\mathcal{Q} H_K^\top}{\sqrt{d_k}}) H_K \quad (5)$$

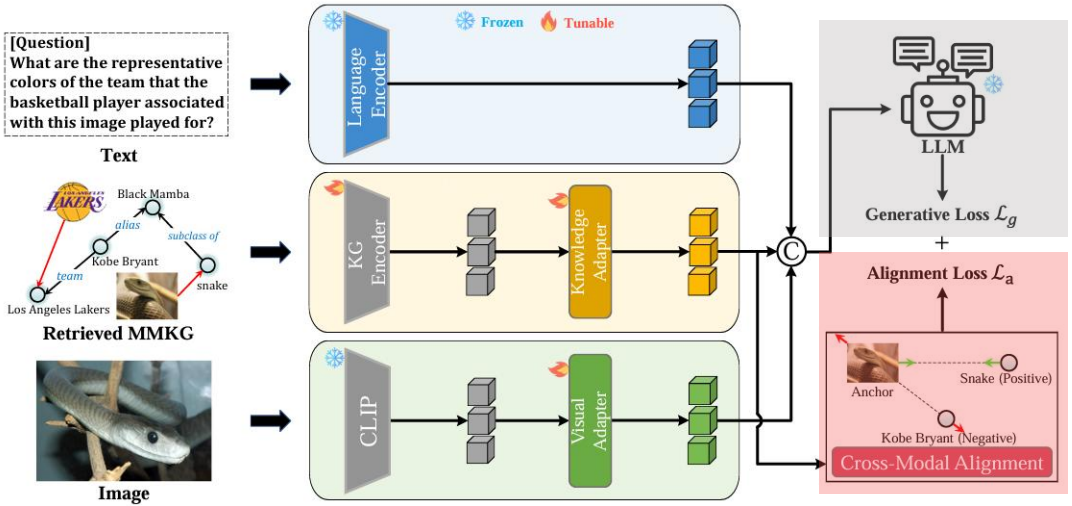
where W_K represents the trainable knowledge adapter matrix, and $\mathcal{Q}$ corresponds to either H_T or H_I, based on the specific scenario at hand.

Ishiwatari, Taichi, et al. "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations." EMNLP 2020.

2 Method

Loss function

- LLM이 auto-regressive하게 답변을 생성할 수 있도록,Causal Language Modeling (CLM) loss를 사용함
- MMKG의 이미지 노드 임베딩이 충분한 semantic 정보를 갖기 위해, Triplet loss를 활용함
- MMKG에 포함된 이미지는 CLIP이 사전에 학습하지 않은 경우가 많기 때문에, 텍스트 임베딩과의 정렬이 필요함



$$\mathcal{L}_g = \sum_{i=1}^{L} \log p(A_i | prompt, A_{0:i-1}; \theta_a) \quad (7)$$

$$\mathcal{L}_a = \sum_{i=1}^{M} \max(d(x_a, x_p) - d(x_a, x_n) + \alpha, 0) \quad (6)$$

where d represents the Euclidean distance, M is the number of selected image entities, and α is a constant used to ensure a certain margin between the distances of positive and negative examples.

2 Method

Knowledge Retrieve Schemes

- 입력된 텍스트 또는 이미지를 기반으로, 관련된 MMKG 서브그래프를 추출하여 입력으로 활용함
- 전체 MMKG에서 query와 유사한 상위 n개의 triple을 1차 필터링함
- 추출된 triple에서 등장한 entity를 중심으로 서브그래프를 제작하고, 다시 query와 관련된 triple을 Top-N으로 선택함 (re-ranking)

Given a query q (text or image) and a multimodal knowledge graph $\mathcal{G}$ with triples $\{(h_i, r_i, t_i)\}_{i=1}^{M}$:

1. **Embedding**:

$$\mathbf{q} = \mathrm{Enc}(q), \quad \mathbf{T}_i = \mathrm{Enc}(h_i, r_i, t_i) \quad \forall i \in [1, M]$$

2. **Similarity Computation**:

$$s_i = \cos(\mathbf{q}, \mathbf{T}_i) \quad \forall i \in [1, M]$$

3. **Top-n Triple Selection**:

$$\mathcal{T}_n = \mathrm{TopK}(\{(h_i, r_i, t_i)\}, \{s_i\}, n)$$

4. **Entity Set Extraction**:

$$\mathcal{E}' = \{h \cup t \mid (h, r, t) \in \mathcal{T}_n\}$$

5. **Subgraph Retrieval**:

$$\mathcal{G}' = \mathrm{Subgraph}(\mathcal{G}, \mathcal{E}', \text{one-hop})$$

6. **Re-ranking in Subgraph**:

$$\mathbf{T}'_j = \mathrm{Enc}(h_j, r_j, t_j), \quad s'_j = \cos(\mathbf{q}, \mathbf{T}'_j) \quad \forall (h_j, r_j, t_j) \in \mathcal{G}'$$

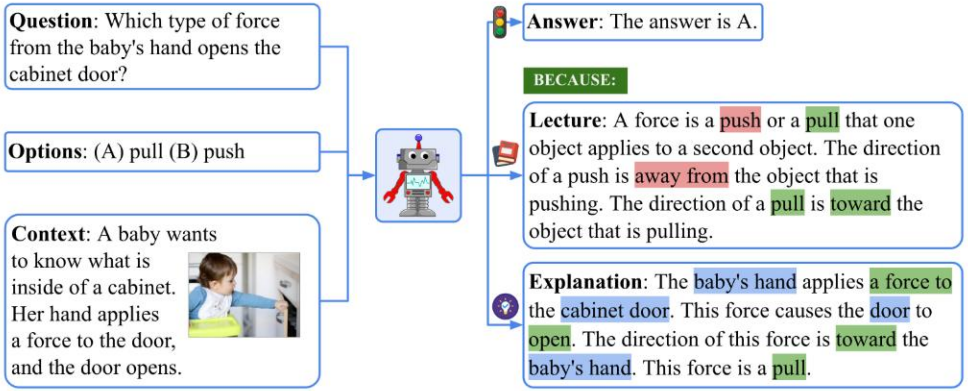
7. **Top-N Final Triple Selection**:

$$\mathcal{T}_N = \mathrm{TopK}(\mathcal{G}', \{s'_j\}, N)$$

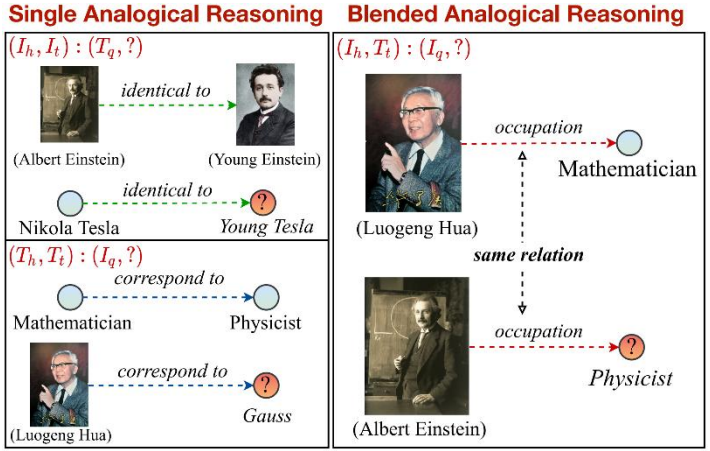
3 Experiments

Datasets

- **ScienceQA:** 대용량의 과학 관련 질문-답변 데이터셋으로, 48.7%의 데이터가 이미지 질문을 포함
- **MARS:** 유사한 관계 패턴을 추론하는 유추 추론 데이터셋으로, 객체는 텍스트 혹은 이미지로 제시됨
- 각각의 데이터셋에 적합한 MMKG와 MarKG라는 지식 그래프를 이용함



Question Answering



Analogical Reasoning

3 Experiments

Qualitative Results

- **MARS:** KG를 사용하지 않으면 입력 이미지에 포함된 H_2O를 보고 water라고 예측하지만, 제안 방식은 MMKG에서 관련된 텍스트를 찾아 올바르게 예측함
- **ScienceQA:** KG를 사용하지 않으면 주(state)의 모양을 인식하지 못해 잘못된 예측을 하지만, 제안 방식은 MMKG에서 관련된 이미지를 찾아 올바르게 예측함

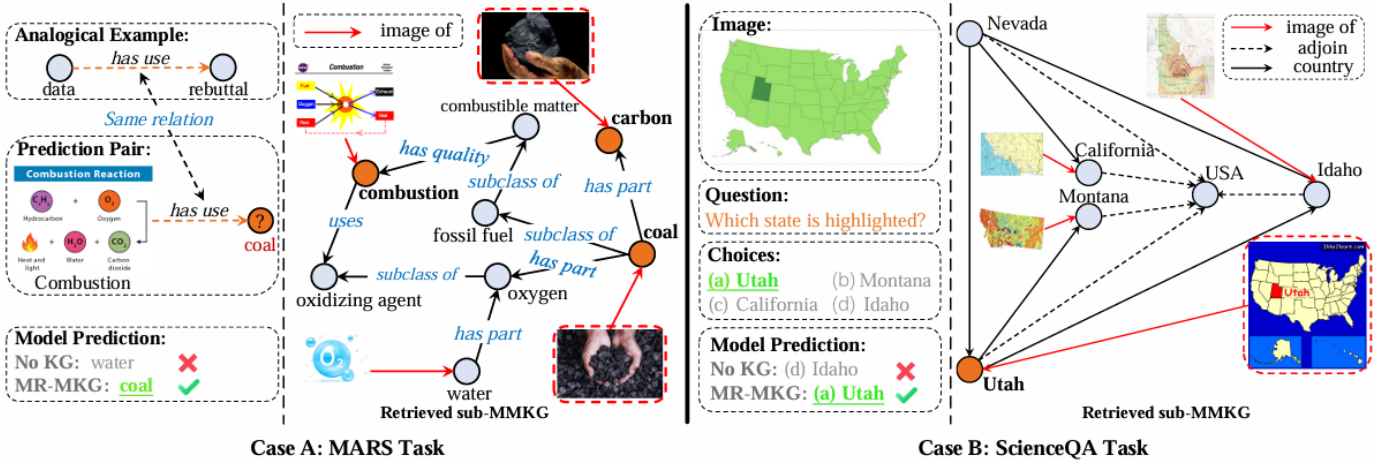


Figure 5: Two examples from MRAS and scienceQA datasets. In case A, the model needs to predict **coal** based on an Analogical Example and the image of combustion. In case B, the model needs to select the correct answer based on the image and the question. Relevant entities for reasoning are marked in orange or highlighted with a red box.

3 Experiments

Comparison with state of the art

- **ScienceQA:** closed-source, open-source, adapter based 방법들과 비교하여 평균적으로 가장 높은 Accuracy를 달성함
- **MARS:** LLM을 사용하지 않은 전통적인 방법들 대비 성능이 높고, LLM을 사용하지만 MMKG를 사용하지 않은 LLaMA 대비 성능이 높음

Method	#T-Param	Subject			Context Modality			Grade		Average
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
Human (Lu et al., 2022)	-	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 (CoT) (Lu et al., 2022)	-	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
GPT-4 (Liu et al., 2023)	-	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
UnifiedQA$_{Base}$ (Lu et al., 2022)	223M	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82	74.11
UnifiedQA$_{Base}$(MM-CoT) (Zhang et al., 2023c)	223M	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
UnifiedQA$_{Large}$(MM-CoT) (Zhang et al., 2023c)	738M	**95.91**	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
LLaVA (Liu et al., 2023)	13B	90.36	**95.95**	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaMA-Adapter (Zhang et al., 2024a)	1.8M	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
LaVIN-7B (Luo et al., 2023)	3.8M	89.25	94.94	85.24	88.51	87.46	88.08	90.16	88.07	89.41
LaVIN-13B (Luo et al., 2023)	5.4M	89.88	94.49	89.82	88.95	87.61	91.85	91.45	89.72	90.83
MR-MKG (FLAN-T5-3B)	77M	90.67	85.38	86.45	90.96	87.46	87.39	90.27	85.23	88.47
MR-MKG (FLAN-T5-11B)	248M	94.93	90.1	90.55	94.53	92.12	92.2	93.83	90.9	92.78
MR-MKG (FLAN-UL2-19B)	248M	95.74	90.33	**92.00**	**95.50**	**92.41**	**93.31**	**93.98**	**93.01**	**93.63**

Table 1: Results on the ScienceQA *test* set with accuracy (%). #T-Params = number of trainable parameters. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12. Previous SOTA results are underlined. The second segment: Zero- & few-shot methods. The third segment: SOTA and representative models. The fourth segment: Parameter-efficient methods. The fifth segment: Our MR-MKG results.

Method	Hits@1	Hits@3	Hits@5	Hits@10	MRR
IKRL (Xie et al., 2017)	0.266	0.294	0.301	0.310	0.283
TransAE (Wang et al., 2019)	0.261	0.285	0.289	0.293	0.276
RSME (Wang et al., 2021)	0.266	0.298	0.307	0.311	0.285
MarT_VisualBERT (Li et al., 2019)	0.261	0.292	0.308	0.321	0.284
MarT_ViLT (Kim et al., 2021)	0.245	0.275	0.287	0.303	0.266
MarT_ViLBERT (Lu et al., 2019)	0.256	0.312	0.327	0.347	0.292
MarT_FLAVA (Singh et al., 2022)	0.264	0.303	0.309	0.319	0.288
MarT_MKGformer (Chen et al., 2022b)	0.301	0.367	0.380	0.408	0.341
Visual_LLaMA-2 7B	0.286	0.373	0.409	0.457	0.347
MR-MKG (Visual_LLaMA-2 7B)	**0.405**	**0.465**	**0.497**	**0.531**	**0.449**

Table 2: Results on the MARS *test* set. The second segment: multimodal knowledge graph embedding (MKGE) methods. The third segment: multimodal pre-trained Transformer (MPT) methods. The fourth segment: MR-MKG. MarT indicates that models are pre-trained on MarKG. Visual_LLaMA means that LLaMA is equipped with a visual adapter.

3 Experiments

Ablation study

- MMKG를 활용하여 Cross-Modal Alignment를 학습하는 방법이 가장 효과적임
- 이미지 질문들만 포함된 ScienceQA 데이터셋에서 성능 향상 폭이 더 큼을 확인할 수 있음
- QA 데이터셋의 특성상 많은 지식이 텍스트에 포함되어 있으므로, 텍스트만을 활용하여 서브그래프를 검색하는 방식이 가장 효과적임

Settings	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	Average
Visual_FLAN-T5-11B	88.45	81.89	84.09	88.47	86.51	85.51	86.75	84.64	86.08(+0.00)
+ KG	93.78	88.64	89.55	93.35	90.47	91.08	92.77	89.65	91.74(+5.66)
+ MMKG	94.23	89.20	90.00	93.94	91.77	91.43	93.39	89.85	92.21(+6.13)
+ Alignment	94.40	89.54	90.09	94.18	91.98	91.50	93.32	90.38	92.36(+6.28)
+ Pre-training	94.93	90.10	90.55	94.53	92.12	92.20	93.83	90.90	92.78(+6.70)

Table 3: Ablation study on the ScienceQA *test* set. "MMKG" indicates using MMKG to replace KG.

Settings	Accuracy (%) on samples
Visual_FLAN-T5-11B	86.59(+0.00)
+ KG	90.37(+3.78)
+ MMKG	91.78(+5.19)
+ Alignment	92.32(+5.73)

Table 4: Ablation study on the samlpes.

Settings	Hits@1 on MARS
Visual_LLaMA-2 7B	0.286(+0.000)
+ KG	0.352(+0.066)
+ MMKG	0.381(+0.095)
+ Alignment	0.394(+0.108)

Table 5: Ablation study on MARS *test* set.

LLM	Method	ScienceQA
FLAN-T5-11B	Text-Only	92.78
	Image-Only	91.58
	Text + Image	92.03

Table 6: Average Accuracy(%) with different subgraph retrieve methods on the ScienceQA *test* set.

Model	Design	ScienceQA	MARS
FLAN-T5-11B /LLaMA-2 7B	GNN	92.23	39.1
	GAT	91.94	39.6
	RGAT	92.78	40.5

Table 7: Impact of different KGE architectures. The metric is Average Accuracy and Hits@1, respectively.

4 Limitations

Limitations

- 부정확한 MMKG의 검색은 LLM의 정확도를 저하시킬 수 있음
- 질문에 맞는 정밀한 지식을 찾아주는 검색 전략 개선이 핵심 연구 방향임
- 또한 작은 LLM과 두 개의 task로 검증하였기에, 대형 LLM을 활용하여 실험을 검증할 필요가 있음

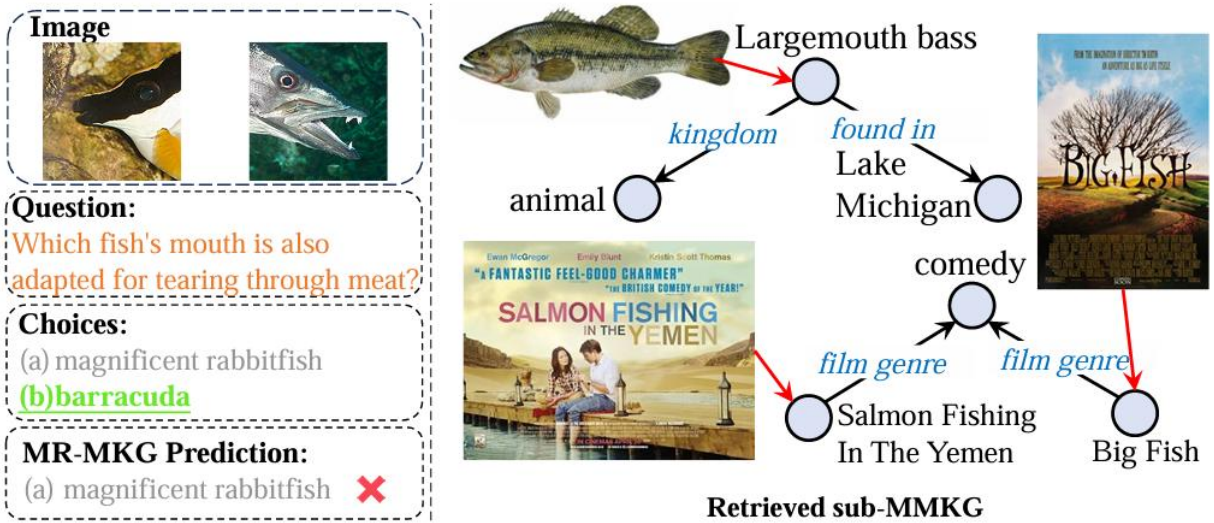
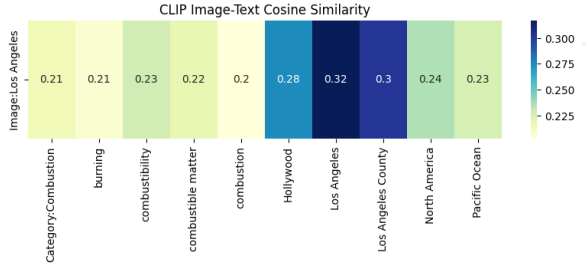


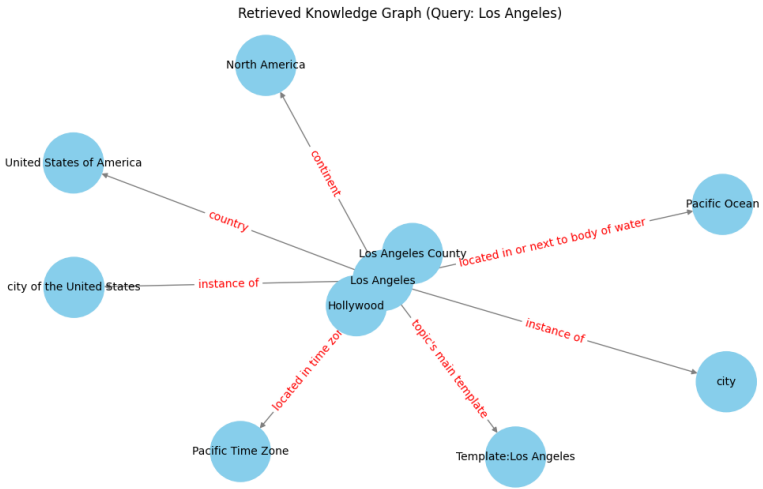
Figure 9: Error example.

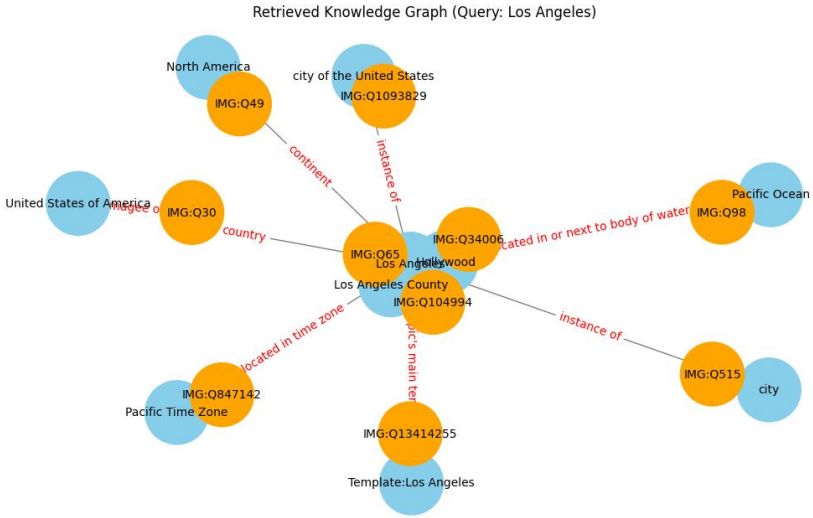
5 Discussion



Knowledge Retrieval

- MARS 데이터셋을 활용하여 MMKG를 구성하고, 서브 그래프를 검색하는 방법을 구현함
- 검색의 효율성을 높이기 위해 KG를 활용하여 서브 그래프를 추출하고, 이미지 노드를 추가하는 방법 사용



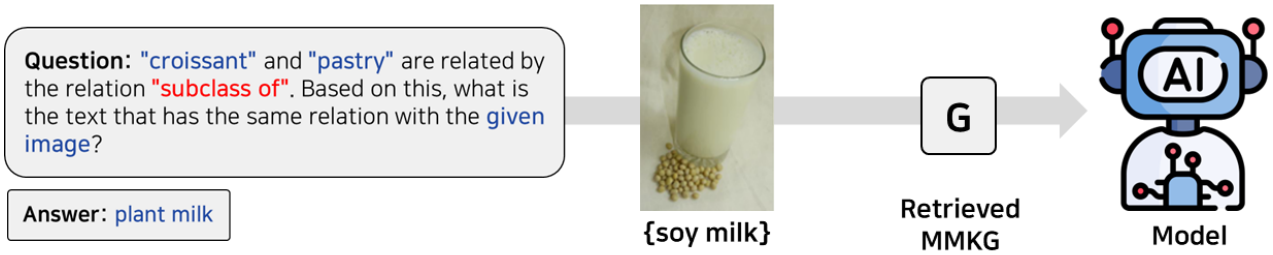


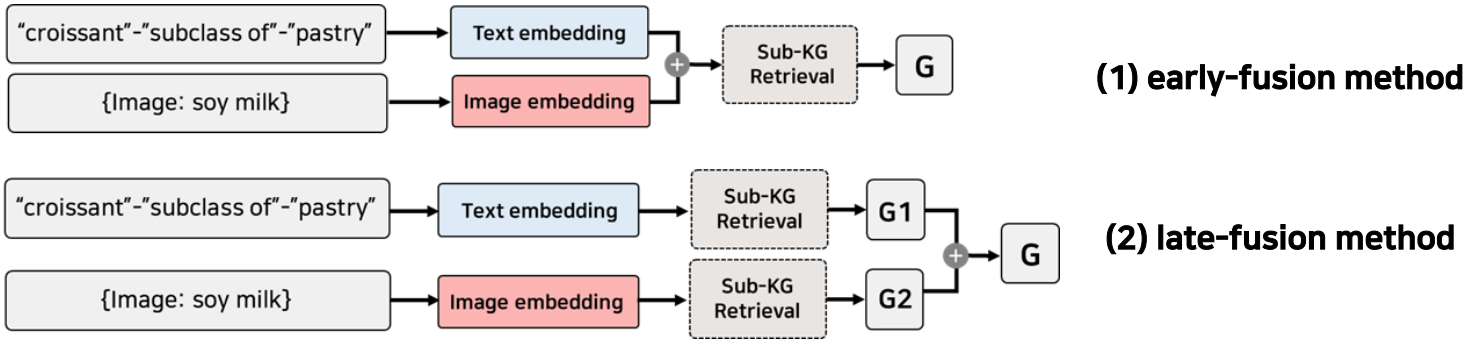
https://github.com/SkiddieAhn/Code-MR-MKG

5 Discussion

Knowledge Retrieval

- Analogical Reasoning을 수행하기 위해, Sub-KG를 검색하는 과정에서 두 가지 방법이 존재할 수 있음
- 첫 번째는 text와 image를 결합하여 query로 활용하는 방식, 두 번째는 각각을 query로 활용하고 검색된 서브 그래프를 결합하는 방식

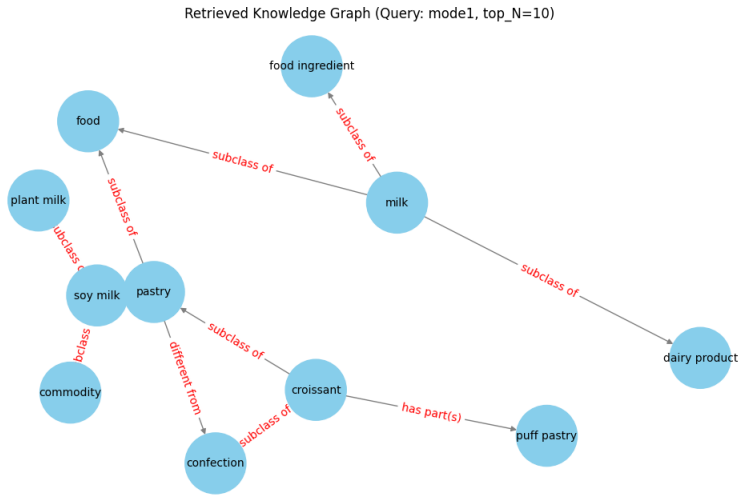




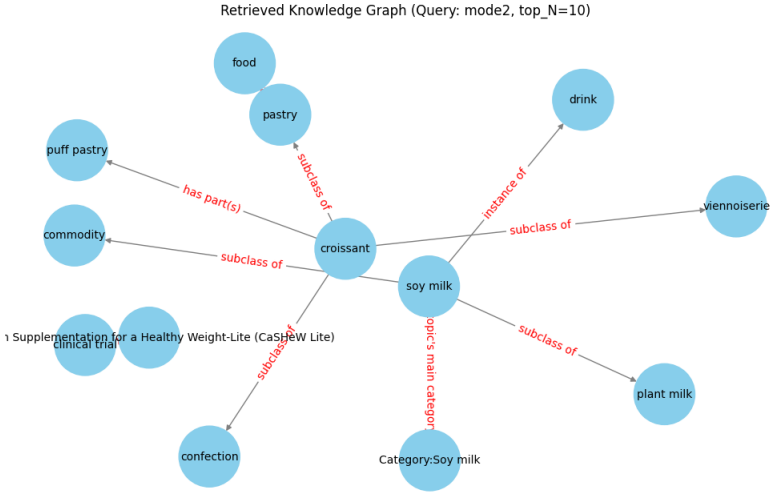
5 Discussion

Knowledge Retrieval

Top-N을 10으로 설정하면 두 방법 모두 핵심 키워드 (soy milk, subclass of, plant milk)가 포함된 것을 확인할 수 있음



(1) early-fusion method

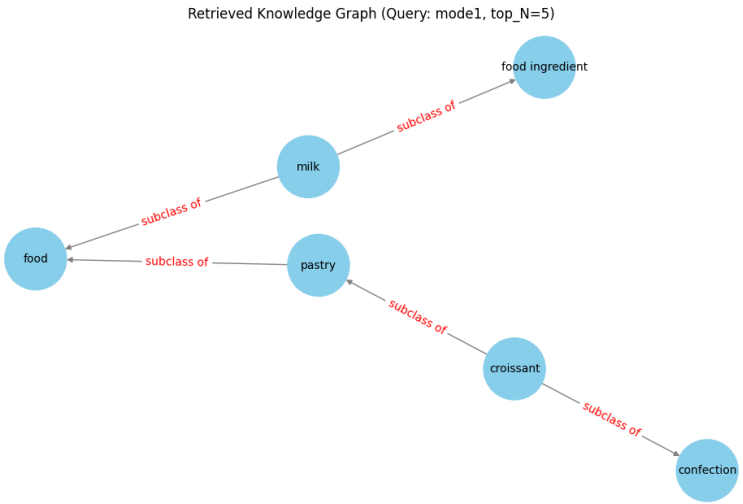


(2) late-fusion method

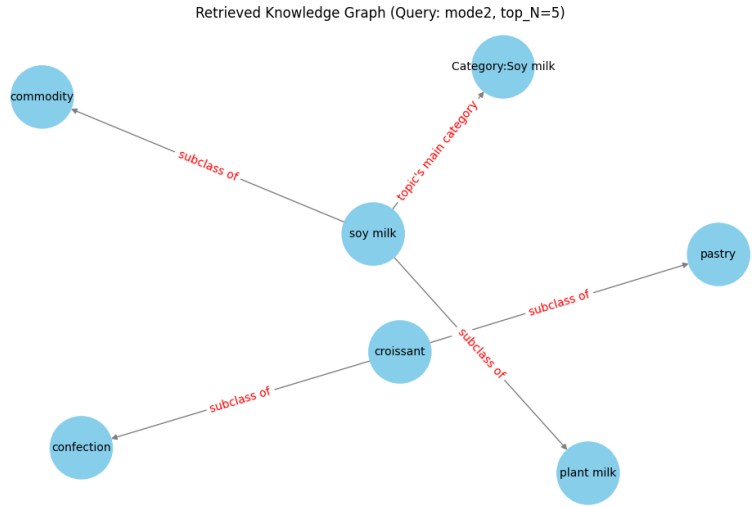
5 Discussion

Knowledge Retrieval

- Top-N을 5로 감소하면 첫 번째 방식에서 핵심 키워드 (soy milk, subclass of, plant milk)가 누락됨을 확인할 수 있음
- 첫 번째 방식은 각 모달이 보유한 고유 정보를 손실시키므로, 올바른 KG 검색이 어려움
- 두 번째 방식은 정상적으로 작동하고 있으나, 이미지-관계를 직접 고려하지 않기 때문에 연결된 엣지가 많은 상황에서는 관계를 제대로 찾지 못할 수 있음



(1) early-fusion method



(2) late-fusion method

Thank you