Modality-Aware Integration with Large Language Models for Knowledge-based Visual Question Answering

Anonymous ACL submission

Abstract

Knowledge-based visual question answering (KVQA) has been extensively studied to answer visual questions with external knowledge, e.g., knowledge graphs (KGs). While several attempts have been proposed to leverage large language models (LLMs) as an implicit knowledge source, it remains challenging since LLMs may generate hallucinations. Moreover, multiple knowledge sources, e.g., images, KGs and LLMs, cannot be readily aligned for complex scenarios. To tackle these, we present a novel modality-aware integration with LLMs for KVQA (`MAIL`). It carefully leverages multimodal knowledge for both image understanding and knowledge reasoning. Specifically, (i) we propose a two-stage prompting strategy with LLMs to densely embody the image into a *scene graph* with detailed visual features; (ii) We construct a coupled *concept graph* by linking the mentioned entities with external facts. (iii) A tailored pseudo-siamese graph medium fusion is designed for sufficient multimodal fusion. We utilize the shared mentioned entities in two graphs as mediums to bridge a tight inter-modal exchange, while maximally preserving insightful intra-modal learning by constraining the fusion within mediums. Extensive experiments on two benchmark datasets show the superiority of `MAIL` with 24× less resources.

1 Introduction

Knowledge-based visual question answering (KVQA) aims to provide appropriate answers for questions about images based on external knowledge (Wang et al., 2017), such as knowledge graphs (KGs) (Marino et al., 2019). It has various applications, especially for assisting the visually impaired users (Gurari et al., 2018), yet still, a challenging task that requires complex reasoning across different data modalities (Yu et al., 2020, 2017).

Recently, several studies have explored using large language models (LLMs) as supplementary knowledge bases and reasoning tools for KVQA (Yang et al., 2022; Gui et al., 2022; Lin et al., 2022); according to how they fuse the knowledge, they can be broadly categorized into direct prompting and modality-agnostic approaches, shown in Figure 1 (a) and (b), respectively. The former directly prompts the question and the corresponding image caption to LLMs for answers (Yang et al., 2022). The latter leverages LLMs to generate candidate answers with supporting evidence and simply combines both question and the external knowledge embedding, e.g., Wikidata (Vrandečić and Krötzsch, 2014), for reasoning at the final stage (Gui et al., 2022; Lin et al., 2022).

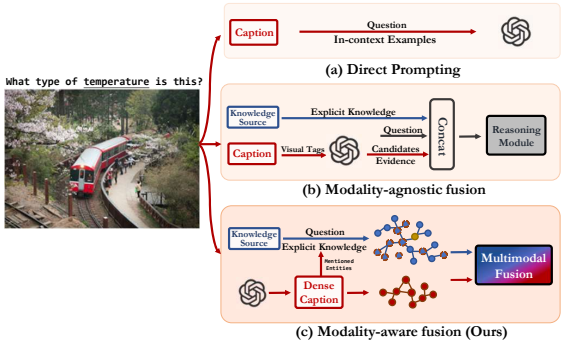

Figure 1: A sketched comparison on employing LLMs for KVQA between existing learning paradigms and ours.

While the above methods have employed LLMs in various ways for KVQA, we argue that they have not fully leveraged the knowledge from LLMs and lack the cross-modal reasoning ability, potentially resulting in sub-optimal performance for complex VQA scenarios. (i) LLMs could incorrectly answer questions or provide unreliable evidence for reasoning. On the one hand, direct prompting to LLMs may struggle to identify the right answer for many complex or domain-specific questions, due to the lack of domain knowledge (Amaro et al., 2023; Shen et al., 2023). On the other hand, LLMs may be prone to generating hallucination (Gravel et al., 2023; Bang et al., 2023) and producing misleading evidence in support of candidate an-

swers. (ii) Integrating multimodal knowledge in a modality-agnostic manner can be sub-optimal. Specifically, existing methods simply concatenate different modal representations, e.g., questions, captions, tags, and external knowledge, for reasoning. This design lacks the necessary cross-modal exchange to enrich the semantics of entities, limiting the final reasoning performance. For example, to correctly answer the question in Figure 1, the model is required to infer the season based on a cross-modal understanding of the inputs, such as the "keep warm" purpose of "coat" and the "spring blooming" feature of "sakura".

In this work, we study the following research question: ***How can we effectively leverage the knowledge from LLMs to enhance the comprehensive understanding and reasoning of the images and questions in KVQA?*** Answering this question is nontrivial due to the following challenges. (i) It is hard to properly incorporate the knowledge from LLMs. LLMs may generate hallucinations when dealing with requests that are not covered in their training corpus. Simply prompting them may generate noisy and irrelevant responses. (ii) Semantic alignment of multiple knowledge sources is challenging. Given image captions, object/region features, external knowledge from KGs, and implicit knowledge from LLMs, appropriately aligning relevant semantic information in different modalities cannot be readily achieved.

To tackle these challenges, we present a novel modality-aware framework to effectively integrate LLMs for KVQA in Figure 1 (c), dubbed `MAIL`. Specifically, (i) we propose a two-stage prompting strategy to maximally leverage the knowledge from LLMs for image understanding. We initialize a dense caption by prompting a visual LLM, e.g., Visual ChatGPT (Wu et al., 2023) and MiniGPT-4 (Zhu et al., 2023). To depict the detailed visual scenes in the caption, we construct a *scene graph* by defining twelve condensed relations and prompting the LLM to extract spatial and object features accordingly in the form of triples, e.g., (*sakura*, *at_location*, *tree*). (ii) We integrate the external knowledge from KGs to form a coupled *concept graph*, where the mentioned entities in scene graphs are linked with real-world assertions and facts to facilitate knowledgeable reasoning, such as (*coat*, *used_for*, *keep warm*) and (*sakura*, *typle_of*, *spring blooming*). (iii) A tailored pseudo-siamese graph medium fusion is designed for effective multimodal graph fusion. Inspired by the success of pseudo-siamese network in measuring the similarity of two correlative inputs (Xia et al., 2021; Gupta et al., 2023), we extend it to graphs to process intra-modal information. It consists of two graph attention networks with the same architecture but different weights. In each sub-encoder, we concentrate on one modality and design a tailored context-aware propagation. This guides our model to attentively prioritize the most valuable entities subject to the particular question. Then we leverage the shared mentioned entities in both coupled graphs as mediums to bridge the cross-modal interaction. The model continuously exchanges their embeddings between two modalities, bringing sufficient complementary knowledge to the other modality respectively. It merely allows inter-modal exchanging by constraining it within the mediums. In general, `MAIL` effectively enhances a tight inter-modal fusion while maximally preserving the insightful intra-modal information for each modality.

Our major contributions are summarized below:

- We formally define a novel learning paradigm, modality-aware integration with LLMs for knowledge-based visual question answering.
- The implicit knowledge in LLMs is carefully leveraged via an effective prompting strategy for coupled scene/concept graph construction.
- We further propose a tailored pseudo-siamese graph medium fusion to integrate multimodal knowledge sources. It balances both intra-modal processing and inter-modal exchange.
- Extensive experiments are conducted on two benchmark datasets. `MAIL` significantly achieves superior performance over a variety of state-of-the-art baselines with $24\times$ less computational resources and $2 \sim 4\times$ faster inferential time.

2 Problem Statement

KVQA requires the model to provide answers to the question $\mathcal{Q}$ of the corresponding image $\mathcal{I}$ based on external knowledge $\mathcal{G}$. In this paper, we propose a novel learning paradigm for leveraging LLMs $f(\cdot)$ for comprehensive knowledge-based VQA.

> Given an image $\mathcal{I}$, a relevant question $\mathcal{Q}$ and external knowledge $\mathcal{G}$, we aim to integrate a visual LLM $f(\cdot)$ and fuse $\{f(\mathcal{I}), \mathcal{Q}, \mathcal{G}\}$ for prediction. The overall performance is evaluated by the accuracy of returned answers with the ground truths.

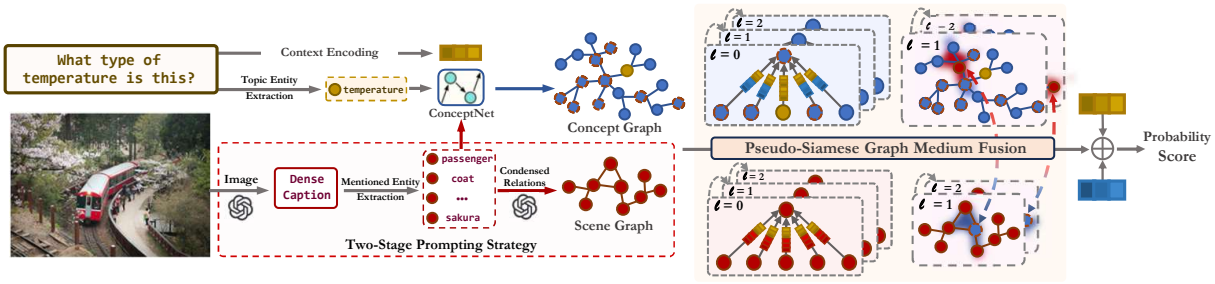

Figure 2: Our proposed framework MAIL, a novel modality-aware integration for knowledge-based VQA with LLMs. Nodes in ***blue*** stand for external knowledge, while ***red*** is for visual objects and ***yellow*** shows the topic entities from questions. Blue nodes with red dashed borders indicate the extracted mediums in concept graph. MAIL is trained to integrate multimodal information for comprehensive cross-modal reasoning with a tailored *PS-GMF*.

3 Methodology

In this section, we introduce the detailed rationale of our proposed framework. An illustration of MAIL is shown in Figure 2. We first carefully leverage the knowledge from LLMs for coupled graph construction. Then, we formulate the pseudo-siamese graph medium fusion (PS-GMF). Through an effective integration of two tailored training objectives, we jointly optimize the model for accurate prediction.

3.1 Scene Graph Construction

Dense Caption Generation We carefully design a hard prompt that requires a visual LLM $f(\cdot)$ to depict the detailed appearance of all the objects in the image and the spatial relations between them. We obtain the generated caption through

$$\mathcal{D} = f(\mathcal{I}, Prompt). \quad (1)$$

We consider the identified visual entities in the image as key mentioned entities appearing in the caption, denoted as $\mathcal{M} = [m_1, m_2 ... m_n] \in \mathcal{D}$. They significantly dominate the multimodal information of both visual features and external knowledge required to answer the questions.

Prompt-enhanced Triple Extraction Given the extracted mentioned entities, we employ LLMs to extract triples. To fully leverage LLMs' comprehension of image captions and prioritize the important visual features, we pre-define 12 relations $\mathcal{R} = [r_1, r_2, ...r_{12}]$ from two aspects: (i) Spatial features. We constrain the description with *at_location*, *next_to*, *in_front_of*, *surrounded_by*, *covered_by*, *includes* and *holds*. (ii) Object features are preserved with not only visual outlooks, i.e., *has_property*, *has_color*, *made_of* and *wears*, but also the intentions of the object if he/she is a human, i.e., *intends_to*. We design a hard template to prompt LLMs for scene graph construction as,

$$\mathcal{G}^S = f(Prompt, \mathcal{D}, \mathcal{M}, \mathcal{R}). \quad (2)$$

We show the detailed statistics and beautiful distributions of all twelve condensed relations in both benchmarks OK-VQA (Marino et al., 2019) and FVQA (Wang et al., 2017) in **Appendix Table 9**.

3.2 Concept Graph Construction

In parallel, we incorporate ConceptNet (Speer et al., 2018) for external commonsense knowledge to construct a concept graph. It is one of the largest knowledge graphs that provides a myriad of structured triples and contains more than eight million real-world entities. We link each mentioned entity m and the topic entity in the question with ConceptNet, and denote the constructed graph as $\mathcal{G}^C$ with sufficient textual descriptions, attributes, categories, and properties of $\mathcal{M}$, that are not present in the image so as to facilitate a more knowledgeable reasoning background for various questions.

3.3 Pseudo-siamese Graph Medium Fusion

Typical pseudo-siamese networks (PSNs) could effectively measure the similarity between two inputs (Gupta et al., 2023; Xia et al., 2021). We extend it to graphs, which naturally fit the requirement of learning coupled graphs for intra-modal processing, leading to pseudo-siamese graph neural networks (PSGs). However, PSG is incapable of cross-modal fusion. Particularly equipped for PSG to enable inter-modal learning, we further design a *graph medium fusion* (GMF) algorithm.

Pseudo-siamese Graph Neural Network Locating valuable entities in different modalities is essential for KVQA. Here, we instantiate PSG with a novel context-aware message propagation scheme to prioritize the most important knowledge in each modality subject to the question context.

Definition. **[Pseudo-siamese GNN]** *We refer to*

PSG Architecture	Formulated Definition
Context-aware Attention	$\mathbf{\Phi}\left(\boldsymbol{m}_t \| \mathbf{c}\right)$
Aggregation Function	$\sum_{t \in \mathcal{N}_h} \alpha_{\boldsymbol{m}_t} \times \boldsymbol{m}_t$
Combination Function	$J(\boldsymbol{e}^{\ell}_{\mathcal{N}_h}) + \boldsymbol{e}^{\ell}_h$
Activation Function	$\begin{cases} \boldsymbol{e}_h, & \textit{if } \boldsymbol{e}_h \geq 0, \\ (1e-2) \times \boldsymbol{e}_h, & \textit{otherwise.} \end{cases}$

Table 1: Formulated definitions of the shared architectures for two sub-networks in the proposed Pseudo-Siamese Graph Neural Network.

a pseudo-siamese graph neural network that consists of two identical graph neural networks for two relevant inputs. They share the same architecture, i.e., attention mechanism, aggregation function, combination function and activation function, but different weights.

As two sub-networks in PSG share the same architecture, we uniformly provide formulations for the intra-modality processing. For each head entity h, we aggregate all the messages from its neighbor tail entities, this set of neighbors is denoted as $\mathcal{N}_h$ and $t \in \mathcal{N}_h$. Since relations in multimodal graphs contain indispensable information for reasoning various real-world questions, we establish the message passing at the triple level, i.e., (h, r, t) to capture abundant semantics as follows.

$$\boldsymbol{m}_{t \in \mathcal{N}_h} = \boldsymbol{W}(\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t), \tag{3}$$

where $(\boldsymbol{e}_h, \boldsymbol{e}_r, \boldsymbol{e}_t)$ is the triple embedding associated with (h, r, t), and $\boldsymbol{W}$ is a learnable matrix for linear transformation. We initialize the entity and relation embedding with a pre-trained language model RoBERTa-large (Liu et al., 2019).

While multimodal graphs always contain desperate information with each other, uniformly training each subnetwork in PSG based on the final prediction lacks awareness of the multimodal characteristics. To this end, we design tailored graph attention networks (Veličković et al., 2017) that allocate a *context-aware* weight $\hat{a}$ to each message, only prioritizing the multimodal messages in both coupled graphs that are highly related to the question. The context-aware weight $\hat{a}_{\boldsymbol{m}_t}$ for each message $\boldsymbol{m}_t$ is correspondingly computed as:

$$\hat{a}_{(h,r,t)} = \mathbf{\Phi}\left(\boldsymbol{m}_t \| \mathbf{c}\right), \tag{4}$$

where $\mathbf{\Phi}$ is the adopted activation function, i.e., LeakyReLU. We endow the attention mechanism to be context-aware by concatenating the question context embedding $\boldsymbol{c}$, expressed as $\|$. Notably, we fix the question context embedding $\boldsymbol{c}$ with RoBERTa and only allow it to participate during the attention allocation process.

By normalizing the attention scores obtained previously, we further assign normative values α to each message $\boldsymbol{m}_t$ of (h, r, t):

$$\alpha_{\boldsymbol{m}_t} = \frac{\hat{a}_{(h,r,t)}}{\sum_{(h,r',t') \in \mathcal{N}_h} \hat{a}_{(h,r',t')}}. \tag{5}$$

To this end, with a weighted sum aggregation operator, we are able to acquire the aggregated representation for entity h in the current layer from its neighbors as $\boldsymbol{e}^{\ell}_{\mathcal{N}_h} = \sum_{(h,r,t) \in \mathcal{N}_h} \alpha_{(h,r,t))} \times \boldsymbol{m}^{\ell}_t$, where the layer number in PSG is denoted as ℓ. We summarize the major functions in Table 1. We finalize the overall architecture of PSG for both inputs from scene graph $\mathcal{G}^S$ and concept graph $\mathcal{G}^C$.

$$\begin{aligned} \boldsymbol{e}^{S(\ell+1)}_h &= J\Big(\sum_{(h,r,t) \in \mathcal{N}_h} \alpha_{(h,r,t))} \times \boldsymbol{m}_t\Big) + \boldsymbol{e}^{S(\ell)}_h, \\ \boldsymbol{e}^{C(\ell+1)}_{\hat{h}} &= J\Big(\sum_{(\hat{h},\hat{r},\hat{t}) \in \mathcal{N}_{\hat{h}}} \alpha_{(\hat{h},\hat{r},\hat{t})} \times \boldsymbol{m}_{\hat{t}}\Big) + \boldsymbol{e}^{C(\ell)}_{\hat{h}}, \end{aligned} \tag{6}$$

where J is a multi-layer perception. The model effectively combines the learned neighbor information $\boldsymbol{e}^{\ell}_{\mathcal{N}_h}$ and itself $\boldsymbol{e}^{\ell}_h$ in current layer. We obtain final representations of all the entities when the layer number ℓ reaches the pre-defined target.

Graph Medium Fusion

In this subsection, we aim to fill the gaps of the aforementioned PSG on *inter-modal* learning. However, there is a challenging dilemma centered around striking the right balance between two crucial aspects. On one hand, we want to maximize the *inter-modal* fusion, where multimodal information could collaborate to yield a more insightful and nuanced understanding of the underlying knowledge subject to answering the question. On the other hand, we recognize the necessity of preserving the integrity of *intra-modal* processing. Considering excessive inter-modal fusion could introduce noise from each other, we aim to maintain the distinctive characteristics and valuable insights that each modality inherently holds.

Since the mentioned entities $\mathcal{M} = [m_1, m_2 ... m_n]$ are shared by $\mathcal{G}^S$ and $\mathcal{G}^C$, we consider these entities existing in both coupled graphs as mediums that possess similar embeddings, since they represent the same real-world object though appearing in different modalities. Motivated by this, we design a novel graph medium fusion algorithm that leverages the medium to bridge two modalities. To get rid of the dilemma, we (i) exchange the representations

of mediums $\boldsymbol{e}_m$ within their respective graphs. This allows the model to delicately introduce cross-modal information with their neighbor entities in the respective graphs, i.e., $\boldsymbol{e}_{\mathcal{N}_m}$; (ii) We strictly impose restrictions on the cross-modal exchange to be within the mediums. This gently brings two modalities closer to each other, while maximally maintaining their individualities. The formulated graph medium fusion process between the coupled graphs is written below.

$$e_m^S = \begin{cases} \boldsymbol{e}_m^S, & \textit{if } \ell = 0, \\ \boldsymbol{e}_m^C, & \textit{otherwise}. \end{cases} \quad e_m^C = \begin{cases} \boldsymbol{e}_m^C, & \textit{if } \ell = 0, \\ \boldsymbol{e}_m^S, & \textit{otherwise}. \end{cases} \tag{7}$$

Specifically, we froze the medium embeddings in the first layer to ensure they have initially aggregated important 1-hop neighbor information. Afterward, the embeddings for the same medium are automatically exchanged after message-passing in the current layer. This sequential approach ensures a high-quality exchange of information between modalities, i.e., visual features and external knowledge, while initially preserving the local context within each modality before they engage in cross-modal interactions during the following layers.

3.4 Training Objective

Answer-targeted Inferential Loss

The primary target of our model is to accurately predict the final answer subject to the particular image and question context. We adopt the binary cross-entropy loss to optimize the inferential performance:

$$\mathcal{L}_{Inference} = -log \frac{MLP(\boldsymbol{e}_a + \mathbf{c})}{\sum_{a' \in \mathcal{G}^C} MLP(\boldsymbol{e}_{a'} + \mathbf{c})}, \tag{8}$$

where a is the correct answer and a' is one of all the candidate answers from $\mathcal{G}^C$. We employ $MLP(\boldsymbol{e}_a + \boldsymbol{c})$ to compute the probability of all the candidate entities in $\mathcal{G}^C$ and prioritize the highest one as the final answer.

Maximum Mean Discrepancy loss

Based on the assumption that one medium in two modalities should be similar to the maximum extent, we approximate their similarity by adopting an auxiliary loss, i.e., Maximum Mean Discrepancy (MMD) loss. The basic kernel function is formulated as follows:

$$\mathcal{K}(\boldsymbol{e}_m^S, \boldsymbol{e}_m^C) = \exp\left(-\frac{||\boldsymbol{e}_m^S - \boldsymbol{e}_m^C||^2}{2\sigma^2}\right), \tag{9}$$

where $\mathcal{K}$ represents the kernel function and σ is a hyperparameter controlling the width of the kernel (Steinwart and Scovel, 2012). Given a valid kernel function where $\mathcal{K}(\boldsymbol{e}_m^S, \boldsymbol{e}_m^C) = (\phi(\boldsymbol{e}_m^S) - \phi(\boldsymbol{e}_m^C))$, we denote the corresponding feature mapping function as ϕ. The final MMD loss for cross-modal alignment is demonstrated hereunder,

$$\mathcal{L}_{Medium} = ||\frac{1}{n}\sum_{m \in \mathcal{M}} \phi(\boldsymbol{e}_m^S) - \frac{1}{n}\sum_{m \in \mathcal{M}} \phi(\boldsymbol{e}_m^C)||^2. \tag{10}$$

We aim to minimize this loss to encourage the learned representations for the same medium from two modalities to be similar in the shared PSG architecture. This effectively guides the process of graph medium fusion by constraining the similarity of mediums in different modalities with each other.

3.4.1 Joint Optimization

The overall framework is jointly optimized according to training objectives as aforementioned. Despite the effectiveness of $\mathcal{L}_{Medium}$, it may introduce inevitable noise by irrespectively forcing the mediums from two modalities to be exactly aligned, which ignores the nature of different modalities. To alleviate this problem, we introduce a hyperparameter λ to control the contribution from $\mathcal{L}_{Medium}$. To this end, the final training loss is calculated below:

$$\mathcal{L}_{Joint} = \mathcal{L}_{Inference} + \lambda \mathcal{L}_{Medium}. \tag{11}$$

4 Experiments

In this section, we conduct a variety of experiments to demonstrate the effectiveness of our proposed `MAIL`. We aim to answer four research questions:

- **RQ1 (Main Results):** How does `MAIL` perform compared with different types of SOTA models?
- **RQ2 (Hyperparameter analysis):** How do hyperparameters influence the performance?
- **RQ3 (Ablation studies):** Does each component eventually contribute to the overall performance?
- **RQ4 (Case study):** How effectively does `MAIL` work in real-world VQA tasks?

4.1 Experimental Setup

Datasets

Following the previous work (Marino et al., 2021; Yang et al., 2022; Gui et al., 2022; Wu et al., 2022; Lin et al., 2022), we mainly conduct our experiments on **OK-VQA** (Marino et al., 2019), which is currently the largest and most challenging benchmark, consisting of 14,055 image-question pairs. To further demonstrate the generalization, we also experimentalize on **FVQA** dataset (Wang et al., 2017), which was the first exploration of KVQA.

Method	Model Inputs	External Knowledge	Fusion Strategy	Acc. (%)
Q Only	Question + Image	-	-	14.93
Traditional End-to-end Baselines				
BAN	Question + Image	-	-	25.17
BAN +AN	Question + Image	Wikipedia	Modality-agnostic	25.61
MUTAN	Question + Image	-	-	26.41
MUTAN +AN	Question + Image	Wikipedia	Modality-agnostic	27.84
ConceptBERT	Question + Image	ConceptNet	Modality-agnostic	33.66
HCNMN	Question + Image	WordNet	Modality-agnostic	36.74
Krisp	Question + Image	Wikipedia + ConceptNet	Modality-agnostic	38.90
MAVEx	Question + Image	Wikipedia + ConceptNet + Google Images	Modality-agnostic	41.37
VLC-BERT	Question + Image	COMET + ConceptNet	Modality-agnostic	43.14
MCAN	Question + Image	-	-	44.65
Large Language Model-enhanced Baselines				
PICa-Base	Question + Caption + Object Tags	Frozen GPT-3 (175B)	-	43.30
Pica-Full	Question + Caption + Object Tags	Frozen GPT-3 (175B)	-	48.00
KAT (Single)	Question + Caption + Object Tags	Frozen GPT-3 (175B) + Wikidata	Modality-agnostic	53.09
KAT (Ensemble)	Question + Caption + Object Tags	Frozen GPT-3 (175B) + Wikidata	Modality-agnostic	54.41
REVIVE	Question + Caption + Region Tags	Frozen GPT-3 (175B) + Wikidata	Modality-agnostic	53.83
MAIL (ours)	Question + Image	Frozen MiniGPT-4 (7B)* + ConceptNet	Modality-aware	**56.69**

Table 2: The overall performance comparison on benchmark dataset OK-VQA. We also elaborate on the detailed comparison with a variety of baselines on the knowledge sources that support their inference, i.e., model inputs, external knowledge, as well as how they fuse multiple modalities.
* We merely leverage it for caption and scene graph construction, with no extra information that is not present in the images.

Method	Fusion Strategy	Acc. (%)
XNM	Modality-agnostic	63.74
KI-Net	Modality-agnostic	63.78
UnifER	Modality-agnostic	66.83
MCAN	-	64.47
HCNMN	Modality-agnostic	69.43
MAIL (ours)	Modality-aware	**73.95**

Table 3: Performance comparison on FVQA.

ACC.%	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	$\ell = 6$
MAIL	56.41	**56.69**	55.45	54.11	52.80

Table 4: Evaluation on the influences of graph layers in pseudo-siamese graph medium fusion.

ACC.%	$\lambda = 0$	$1e-5$	$1e-4$	$1e-3$	$1e-2$	$1e-1$
MAIL	53.34	54.18	55.31	**56.69**	54.30	55.82

Table 5: Exploring the control over the impacts from $\mathcal{L}_{Medium}$ to preserve insightful intra-modal learning.

Baselines

We adopt two pipelines of off-the-shelf methods for performance comparison. Details are demonstrated in the **Appendix A.3**. (i) Traditional end-to-end baselines that design various multimodal learning algorithms for final reasoning over the posed questions. (ii) LLM-enhanced baselines that leverage LLMs, i.e., GPT-3, for direct answer prediction or relevant supporting evidence generation.

4.2 Main Results

To answer **RQ1**, in Table 2 & 3, we summarize the comparisons with all the important baselines. The performance is evaluated by the soft accuracy following previous research (Hu et al., 2023). MAIL outperforms all the traditional baselines regardless of their various knowledge sources and the advantages of leveraging a feature-level image representation. MAIL achieves 12.04% improvements over the best traditional baseline, i.e., MCAN, on OK-VQA and 14.7% on FVQA. For LLM-enhanced baselines, it is worth mentioning that they have utilized the generative ability from (Lin et al., 2022; Brown et al., 2020), which makes them especially advantageous in answering subjective questions, for instance, 'Can people travel on the freeway' or 'Is it illegal?'. Despite this, MAIL still outperforms the best LLM-enhanced baseline with 2.28% increases in general, let alone 13.39% over PICa.

Moreover, MAIL is resource-efficient, requiring the smallest number of parameters among all the LLM-enhanced baselines, shown in Table 6. We have used significantly far fewer parameters than any other LLM-enhanced models, i.e., 7.13 B, for answer prediction. As a result, the inferential time of MAIL for one test question is 0.661s (when batch size = 1). Generally, existing LLM-enhanced baselines commonly utilize over 24 times more parameters and 2∼4 times of inferential time than MAIL.

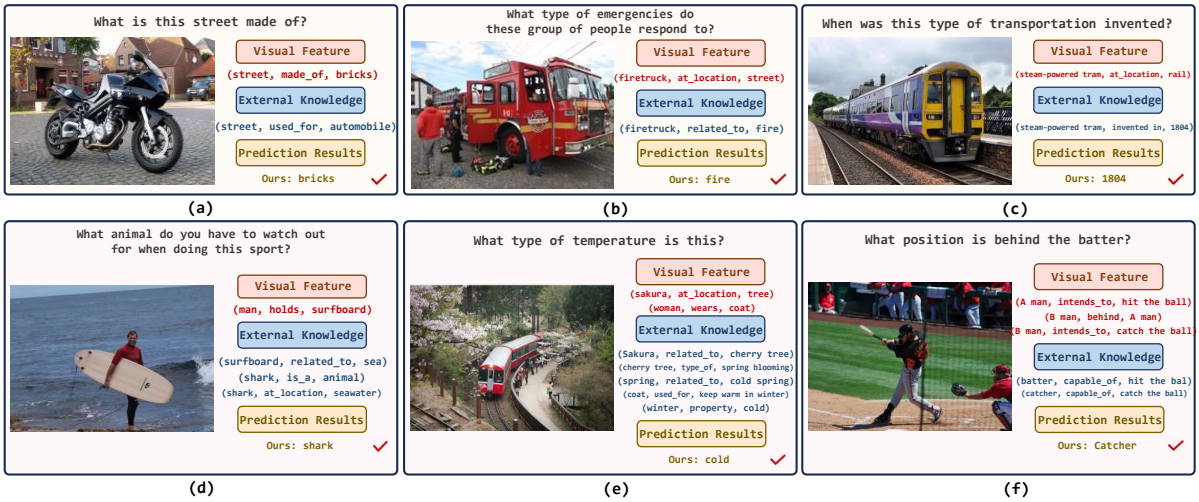

Figure 3: Case studies with both single-hop and multi-hop reasoning examples in OK-VQA.

Models	∼Param Size	Training Time	Inference Time
PICa	∼175.00 B	/	1.547 s
KAT	∼175.80 B	3.025 s	1.292 s
REVIVE	∼175.80 B	4.500 s	2.644 s
MAIL(Ours)	**∼7.13 B**	**2.699 s**	**0.661 s**

Table 6: Comparisons on the computational costs and inferential time with LLM-enhanced baselines.

Reasoning Module	Accuracy (%)
PSG (w/o GMF)	55.53
PS-GMF	**56.69**

Table 7: Verification of the importance of inter-modality fusion by removing GMF with PSG only.

Pure LLMs	Multimodal Understanding	Acc.(%)
Large Language Models		
Llama (7B)	Dense Caption	39.27
Llama2 (7B)	Dense Caption	45.35
ChatGPT (GPT3.5)	Dense Caption	40.26
GPT-4	Dense Caption	54.33
Visual Large Language Models		
Visual ChatGPT	BLIP-VQA-Base + GPT3.5	38.70
MiniGPT-4 (7B)	ViT + Vicuna	51.26
Ours	Dense Caption + PS-GMF	**56.69**

Table 8: Ablation studies on comparing with pure LLMs by directly feeding the questions and (i) corresponding image caption to LLMs or (ii) the raw images to visual LLMs for answers in a zero-shot setting.

4.3 Hyperparameter Analysis

Search of Graph layers

The main architecture of PS-GMF naturally comprises the discussion of the impacts from graph layers ℓ. We empirically hypothesize that augmenting the depth of the ℓ could facilitate both a deeper understanding of single modalities (i.e., PSG) and a more profound exchange of information between modalities (i.e., GMF). However, it remains unclear about when to reach the plateau. Simply adding more layers may over-fuse two modalities and lose the ability of intra-modal processing, while reducing layers may lead to an adverse situation with inadequate inter-modal fusion. To this end, we vary the layer number and show the performance changes in Table 4. The final prediction performance of MAIL is reported when $\ell = 3$.

Investigation on hyperparameter λ

While an excessively strict alignment of mediums may homogenize the intra-modal information, we aim to find a suitable λ that constrains the impacts of $\mathcal{L}_{Medium}$. This could significantly encourage harmonious inter-modal fusion from multiple modalities while retaining the richness and specificity inherent to each modality. The experimentation process involves a systematic adjustment of λ across a range of values, specifically within the interval $[0, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1]$. We showcase the results in Table 5. Upon careful examination of the performance trends, we employ $\lambda = 1e-3$ for a balanced trade-off.

4.4 Ablation Studies

Empirical comparison with LLMs

In this ablation study, we further demonstrate our tailored multimodal learning module PS-GMF, and delineate the specific contributions by comparing it against frozen LLMs. Specifically, we adopt both pure LLMs, i.e., Llama (Touvron et al., 2023a) and Llama2 (Touvron et al., 2023b), as well as visual

LLMs with Visual ChatGPT (Wu et al., 2023) and MiniGPT-4 (Zhu et al., 2023). We exclusively constrain the inputs in a zero-shot setting with only dense captions and questions for LLMs, while raw images and questions for frozen visual LLMs. The results are summarized in Table 8. MAIL outperforms the best LLM GPT-4 with 2.36% improvements, attributed to the effective graph medium fusion that integrates external knowledge. MAIL also significantly outperforms Visual ChatGPT and MiniGPT-4 with 17.99% and 5.43% higher accuracy. The results shed light on the cross-modal reasoning ability of MAIL.

Reasoning with PSG Only

In this subsection, we explore the importance of inter-modality interaction by removing the graph medium fusion and only relying on PSG for inference. We list the performance of 'PSG w/o (GMF)' in Table 7. The complete multimodal reasoning with PS-GMF outperforms the version with only intra-modal learning with 1.16% improvements. Under this PSG-only setting, we seek to grasp insights into the necessity of graph medium fusion for fostering effective inter-modality interaction. Understanding the performance impact of omitting this fusion mechanism supports the value of shared entities and medium exchange in bridging the cross-modal interaction and facilitates our proposed modality-aware integration with LLMs.

4.5 Case Studies

In this section, we answer **RQ4** with six real-world examples from OK-VQA in Figure 3 to shed light on our effectiveness. **Single-hop questions** can be directly inferred with easily accessible information from either the visual content or external knowledge sources, while **multi-hop questions** pose more challenges for accurately locating answers several hops away from mentioned entities.

These cases show the adeptness of MAIL in handling a spectrum of questions, requiring both straightforward inferences from explicit information and complex multi-hop reasoning ability by integrating implicit knowledge sources. For example, Figure 3 (a) can be answered based on the visual information captured by the scene graph without external knowledge, while the answer of (e) needs to be artfully inferred from two different angles, i.e., both the blossom season of sakura and the warmth of people's clothes. These can be attributed to (i) the coupled graph construction that contains abundant modality-aware knowledge to ground the reasoning, as well as (ii) the effective design of our pseudo-siamese graph neural network. It benefits sufficient preservation of intra-modal information and adequate cross-modal fusion, resulting in a powerful multi-hop reasoning ability over both inherent visual features and external knowledge.

5 Related Work

KVQA with KGs. Early studies either dedicated to integrating different knowledge sources (Wang et al., 2017) or proposed various fusion algorithms for multimodal information (Marino et al., 2021). ConceptBERT (Gardères et al., 2020) constrains the multimodal information with question embedding and fuses embeddings of each modality for prediction. MAVEx (Wu et al., 2022) aims to discern the corresponding knowledge source for each candidate answer to reduce noise. KRISP (Marino et al., 2021) captures both implicit information in both questions, images and knowledge graphs.

KVQA with LLMs. Recently, large language models (LLMs) have surprised the community with their superior understanding of texts. PICa (Yang et al., 2022) first leverages GPT3 (Brown et al., 2020) as an implicit knowledge source for reasoning by prompting the image captions and in-context examples. Another pipeline of studies employs LLMs to generate candidates or supporting evidence for particular captions, e.g., KAT (Gui et al., 2022) and REVIVE (Lin et al., 2022). While they do not fully leverage the multiple sources of knowledge, we break the limitation of complex reasoning by developing a tailored multimodal fusion algorithm that balances intra- and inter-modal learning.

6 Conclusions

We present MAIL, a modality-aware integration with large language models for knowledge-based visual question answering. We formally define a novel multimodal learning paradigm for comprehensive cross-modal reasoning among multiple knowledge sources. The knowledge from LLMs is effectively leveraged via a carefully designed coupled graph construction, i.e., scene graph and concept graph. Then we integrate various multimodal information with a tailored pseudo-siamese graph medium fusion. It effectively enhances a tight inter-modal interaction and maximally preserves insightful intra-modal processing. MAIL achieves superiority on two benchmark datasets while possessing 24$\times$ less computational resources and 2$\sim$ 4$\times$ faster inferential time than the existing state-of-the-art baselines.

Limitations

We adopt the popular visual LLM, i.e., MiniGPT-4 (7B) as the knowledge source for image caption generation and scene graph construction. While more advanced visual LLMs have emerged recently, e.g., GPT-4, Gemini Vision Pro, etc, we will further enrich our experiments and comparisons with updated captions and scene graphs as future work.

Ethics Statement

We confirm that we have fully complied with the ACL Ethics Policy in this study. We conduct experiments with widely adopted publicly available datasets. The generated image captions, processed scene graphs and concept graphs will be open-sourced for other researchers' fair reproduction and further study in the active KVQA community.

References

Ilaria Amaro, Attilio Della Greca, Rita Francese, Genoveffa Tortora, and Cesare Tucci. 2023. Ai unreliable answers: A case study on chatgpt. In *ICHCI*, pages 23–40. Springer.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, pages 2612–2620.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS*, 33:1877–1901.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. Conceptbert: Concept-aware representation for visual question answering. In *EMNLP*, pages 489–498.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander G Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL*, pages 956–968.

Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A unified end-to-end retriever-reader framework for knowledge-based vqa. In *ICMM*, pages 2061–2069.

Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. 2023. Siamese masked autoencoders. *CVPR*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided task-aware image captioning. *ICCV*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *NeurIPS*, 31.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *NeuIPS*, 35:10560–10571.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *CVPR*, pages 14111–14121.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. Vlc-bert: visual question answering with contextualized commonsense knowledge. In *WACV*, pages 1155–1165.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8376–8384.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge.

Ingo Steinwart and Clint Scovel. 2012. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *TPAMI*, 40(10):2413–2427.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based vqa. In *AAAI*, volume 36, pages 2712–2721.

Congying Xia, Caiming Xiong, and Philip Yu. 2021. Pseudo siamese network for few-shot intent generation. In *SIGIR*, pages 2005–2009.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, volume 36, pages 3081–3089.

Jing Yu, Zihao Zhu, Yujing Wang, Weifeng Zhang, Yue Hu, and Jianlong Tan. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. *Pattern Recognition*, 108:107563.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *CVPR*, pages 1821–1830.

Yifeng Zhang, Shi Chen, and Qi Zhao. 2023. Toward multi-granularity decision-making: Explicit visual reasoning with hierarchical knowledge. In *ICCV*, pages 2573–2583.

Yifeng Zhang, Ming Jiang, and Qi Zhao. 2021. Explicit knowledge incorporation for visual reasoning. In *ICCV*, pages 1356–1365.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Prompt Templates for Coupled Graph Construction

Prompt for Scene Graph Construction

'Describe the image with as many details as possible. Generally, identify the objects and their spatial relations with each other. Specifically, include the visual outlook of different objects, e.g., color, style as well as the appearance for human beings.'

Prompt for Concept Graph Construction

'Given the image caption, based on your comprehensive understanding, construct a high-quality scene graph with as many meaningful details of the mentioned entities as possible in the form of a triple (head entity, relation, tail entity). $\backslash n$ Strictly use the twelve predefined relations from: $\mathcal{R}$, e.g., (woman, in_front_of, car), (car, has_color, blue), only return the triples with no other content. $\backslash n$ Caption: $\mathcal{D}$ $\backslash n$ Mentioned Entities: $\mathcal{M}$.'

A.2 Detailed Statistics of the Scene Graphs

We showcase the beautiful distribution of the pre-defined condensed relations in the constructed scene graphs for OK-VQA and FVQA in Table 9.

Categories	Relation	OK-VQA		FVQA	
		Tain	Test	Tain	Test
Spatial Features	at_location	10,562	10,118	3,466	3,107
	next_to	3,948	3,772	2,533	2,289
	in_front_of	2,239	2,244	759	687
	surrounded_by	2,004	2,026	699	549
	covered_by	180	191	9	7
	includes	12,402	12,390	1,811	1,630
	holds	3,344	3,090	965	794
Object Features	has_property	16,685	17,032	1,301	1,297
	has_color	9,191	8,836	3,653	3,258
	made_of	3,388	3,310	978	948
	wears	5,172	5,049	1,504	1,449
	intends_to	1,599	1,655	9	8

Table 9: The overall statistics of the pre-defined condense relations for OK-VQA and FVQA datasets. They depict the spatial features and object features in images.

A.3 Experiments

Implementation Details We generate dense image captions with MiniGPT-4 (7B) (Zhu et al., 2023), and adopt ConceptNet (Speer et al., 2018) for external knowledge, one of the largest real-world commonsense KGs. We apply MiniGPT-4 with one Tesla V100. The entire processing of OK-VQA and the corresponding Microsoft COCO images (Lin et al., 2014) including image-to-text and data cleaning takes about 4 rounds. We adopt $\ell = 3$ and $\lambda = 1e-3$ after hyperparameter tuning. The generated caption is stored for further multimodal learning. Our codes and processed graphs will be open-sourced and publicly available.

For the results of baseline LLMs, since they could occasionally refuse to answer with responses like either 'As a language model, I am not capable of understanding images' or 'Sorry, there is no related information in the provided caption.', we report the average accuracy over 2 rounds. **Baselines** Specifically, for traditional end-to-end baselines, we pick the representative state-of-the-art methods, i.e., a direct answering based on questions only (Q Only) (Marino et al., 2019),BAN (Kim et al., 2018), MUTAN (Ben-Younes et al., 2017), ConceptBERT (Gardères et al., 2020), KRISP (Marino et al., 2021), MAVEx (Wu et al., 2022), VLC-BERT (Ravi et al., 2023), HCNMN (Zhang et al., 2023) and MCAN (Yu et al., 2019). Moreover, as BAN and MUTAN merely learn the uni-modal visual features, they are augmented with ArticleNet (AN) (Marino et al., 2019) that is trained to retrieve knowledge from Wikipedia for corresponding question-image pair to facilitate the reasoning with external knowledge, denoted as 'BAN + AN' and 'MUTAN + AN' (Marino et al., 2019).

While for LLM-enhanced baselines, we adopt PICa (Yang et al., 2022), KAT (Gui et al., 2022), and REVIVE (Lin et al., 2022).

A.4 Generalization on FVQA Dataset

To further demonstrate the generalization ability of our proposed `MAIL`, we compare it with the widely adopted baselines on the first KVQA dataset **FVQA**, i.e., XNM (Shi et al., 2019), KI-Net (Zhang et al., 2021), UnifER (Guo et al., 2022), MCAN (Yu et al., 2019) and HCNMN (Zhang et al., 2023). For external knowledge, KI-Net uses ConceptNet and Wordnet; UnifER uses Visual-Bert, LXMERT and ViLT; HCNMN uses WordNet, WikiText, ConceptNet and Visual Genome.