MAGIC-VQA: Multimodal And Grounded Inference with Commonsense Knowledge for Visual Question Answering

Anonymous ACL submission

Abstract

Visual Question Answering (VQA) requires reasoning across visual and textual modalities, yet Large Vision-Language Models (LVLMs) often lack integrated commonsense knowledge, limiting their robustness in real-world scenarios. To address this, we introduce MAGIC-VQA, a novel framework that enhances VQA by systematically integrating commonsense knowledge with LVLMs. MAGIC-VQA employs a three-stage process: (1) Explicit Knowledge Integration from external sources, (2) By-Type Post-Processing for contextual refinement, and (3) Implicit Knowledge Augmentation using a Graph Neural Network (GNN) for structured reasoning. While GNNs bring greater depth to structured inference, they enable superior relational inference beyond LVLMs. MAGIC-VQA bridges a key gap by unifying commonsensse knowledge with LVLM-driven reasoning, eliminating the need for extensive pre-training or complex prompt tuning. Our framework achieves state-of-the-art performance on benchmark datasets, significantly improving commonsense reasoning in VQA.

1 Introduction

Visual Question Answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Yue et al., 2024) is a complex task requiring models to understand the interaction between visual inputs and textual queries. In recent years, Large Vision-Language Models (LVLMs) (Dai et al., 2023; Xue et al., 2024; Wang et al., 2024b; Chen et al., 2023, 2024b; Liu et al., 2024b; Li et al., 2024; OpenAI, 2024b; Gemini Team, 2024) have made substantial progress in VQA through extensive pre-training on massive image-text datasets and instruction tuning. These models excel at object-level visual recognition and semantic understanding, capturing attributes such as spatial relationships and contextual details.

Nevertheless, LVLMs often face challenges on questions requiring commonsense reasoning—particularly those hinging on implicit contextual cues or everyday world knowledge (Zhou et al., 2023; Ye et al.; Li et al., 2023b)[1]. To overcome this limitation and improve the performance on commmonsense VQA, different methods have been explored. For example, Multimodal retrieval-augmented generation leverages dense retrieval to inject external multimodal information into the generation process, thereby enhancing the factual grounding of LVLMs (Lin and Byrne, 2022; Hu et al., 2023). Multimodal prompt tuning harnesses the model's innate commonsense knowledge by carefully crafting prompts that combine visual and textual cues from representative samples, guiding LVLMs to leverage their internal reasoning for context-rich answers (Wei et al., 2022; Zhang et al., 2023). However, static prompt design usually lacks the dynamic adaptability required for novel scenarios, resulting in limited generalization to unseen or diverse inputs. Additionally, graph-based approaches utilize Graph Neural Networks (GNNs) to incorporate structured commonsense knowledge (Ravi et al., 2023; Wang et al., 2022), which surpasses the limitations of purely parametric LVLMs, enabling models to capture explicit and implicit knowledge connections via structured graphs.

However, a key missing component in existing works is the effective integration of commonsense knowledge with LVLMs while addressing their inherent shortcomings. Prior approaches either rely on static retrieval that indiscriminately injects input-unaware noisy knowledge or graph-based augmentation that overlooks the dynamic interplay between external and innate knowledge. Our work seeks to fill this gap by proposing a unified framework that systematically combines dynamic, contextually aligned commonsense integration with structured graph-based reasoning to robustly filter and

[1]The sample illustrations can be found in Section 6 and Appendix.

incorporate relevant commonsense knowledge.

In this paper, we introduce MAGIC-VQA, a novel framework designed to enhance VQA models by effectively integrating commonsense knowledge with LVLMs. MAGIC-VQA is built upon a three-stage process that not only improves reasoning capabilities but also mitigates the complexity associated with large-scale pre-training and inefficient prompt-based approaches. First, explicit Commonsense Knowledge Integration extracts relevant knowledge triples from external sources, establishing a reliable reasoning foundation. Secondly, by-Type Commonsense Knowledge Post-processing refines these triples based on input-specific needs, ensuring contextual relevance. Finally, implicit Commonsense Knowledge Augmentation constructs a heterogeneous multimodal graph processed by a GNN to capture intricate relationships, providing structured reasoning beyond what LVLMs alone can infer. By integrating explicit and implicit commonsense knowledge on top of LVLMs, MAGIC-VQA addresses both the limitations of previous approaches and the missing component in existing works. Our main contributions are as follows:

1. We propose MAGIC-VQA, a novel end-to-end framework that systematically integrates both explicit and implicit commonsense knowledge into VQA through, without extensive pre-training or intricate prompt tuning.

2. MAGIC-VQA employs a three-stage pipeline—explicit commonsense integration, by-type post-processing, and graph-based implicit augmentation-that dynamically extracts and filters commonsense knowledge in an input-aware manner, and leverages a GNN-based structured reasoning mechanism.

3. We conduct extensive evaluations across multiple VQA benchmarks, demonstrating robust improvement in commonsense understanding reasoning for VQA, surpassing existing models in both knowledge grounding and inference accuracy.

2 Related Work

2.1 VLPM and LVLMs on VQA

Vision-Language Pretrained Models (VLPMs) like ViLBERT (Su et al., 2019), ALBEF (Li et al., 2021) and VILT (Kim et al., 2021) have advanced Visual Question Answering (VQA) by improving the alignment between visual and textual modalities in the last few years. Recently, Large Vision-Language Models (LVLMs) like InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2024b),GPT4o (OpenAI, 2024b) and Gemini1.5 (Gemini Team, 2024) further push the boundary of VQA with strong in-context learning capability through extensive pre-training and instruction-tuning on large-scale image-text datasets. However, these models still face challenges with questions requiring commonsense knowledge that is intuitive and straightforward for humans, such as reasoning based on implicit contextual cues or general world knowledge (Ye et al., 2023; Wang et al., 2023; Chen et al., 2024a; Yang et al., 2024). The resource-intensive nature of these models further makes it infeasible to train a model from scratch specifically for enhanced commonsense understanding.

2.2 Commonsense Knowledge Integration for Visual Question Answering

Several studies have highlighted the critical role of commonsense knowledge integration in enhancing the performance of VLPMs and LVLMs on VQA tasks (Wu et al., 2022; Zhang et al., 2022; Ding et al., 2022; Wang et al., 2024c). These methods can be classified into two approaches: explicit commonsense knowledge integration and implicit commonsense knowledge integration.

1) Explicit commonsense knowledge integration directly incorporates external commonsense knowledge into model training through instruction tuning or prompt tuning. For example, VLC-BERT (Ravi et al., 2023) encodes the contextualized commonsense knowledge of the question phrases as additional textual features and integrates with object visual features to fine-tune the VL-BERT (Su et al., 2019). MM-CoT (Zhang et al., 2023), T-SciQ (Wang et al., 2024a) and KAM-CoT (Mondal et al., 2024) fine-tune models on commonsense-augmented Chain-of-Thought (CoT) data to enhance their reasoning processes. However, these methods suffer from static commonsense integration without dynamic filtering to adjust to varying input contexts, resulting in potential noise that impedes nuanced reasoning.

2) Implicit Commonsense Knowledge Integration focuses on distilling knowledge from a teacher to a student model without directly incorporating external datasets. For example, (Dai et al., 2022) distill knowledge from the dual-stream VLP model CLIP (Radford et al., 2021) into BART (Lewis,

KG	Size	Main Coverage	Key Relations
ConceptNet	8M	PE	*IsA, UsedFor*
ATOMIC	877K	EC, SI	*xWant, oEffect*
ATOMIC2020	1.33M	PE, EC, SI	*23 relation types*

Table 1: Comparison of three commonsense knowledge graphs. 'PE' refers to physical entity-related commonsense, 'EC' to event-centered related commonsense, and 'SI' to social interaction-related commonsense.

2019), achieving strong zero-shot performance on VQA. Park et al. (2024) proposed a novel method to distill knowledge from LLMs focusing on specific image regions, then guiding the LLM to infer commonsense knowledge about those areas. These methods often overlook the structured interplay among visual, textual, and commonsense cues, limiting their ability to perform nuanced reasoning.

3 MAGIC-VQA

MAGIC-VQA employs a three-stage process to integrate commonsense knowledge into LVLMs, as in Figure 1. (1) Explicit Commonsense Knowledge Retrieval extracts relevant triples from an external knowledge graph. (2) By-Type Commonsense Knowledge Post-processing refines these triples, aligning them with dataset-specific distributions and assigning relevance levels. (3) Implicit Commonsense Knowledge Augmentation constructs a multimodal graph processed by a GNN to generate confidence scores. These scores, along with the refined triples with relevance level, image, and question, form a comprehensive input to the LVLMs for robust commonsense-grounded inference.

3.1 Explicit Commonsense Knowledge Integration

We begin by integrating explicit commonsense knowledge into LVLM for each input modality. Given a dataset sample consisting of an image I and an associated question Q, we first generate an image caption C using BLIP2 (Li et al., 2023a) as additional contextual information. Next, we encode the inputs $\{I, Q, C\}$ into a shared embedding space, obtaining representations f_I, f_Q, and f_C using the same model.

We adopt ATOMIC2020 (Hwang et al., 2021) as our external knowledge source because of its broad coverage of physical-entity (PE), event-centered (EC), and social-interaction (SI) relations, as shown in Table 1. Spanning 1.33 million triplets and 23 relation types, it offers a more balanced scope than either ConceptNet (Speer et al., 2017) or the earlier ATOMIC (Sap et al., 2019), making it especially relevant for everyday objects, actions, and social contexts encountered in VQA. These 23 relations fall into three groups: (1) **Physical Entity (PE)**: object properties and functions like "paper is made of cellulose". (2) **Event-Centered (EC)**: situational sequences or events, such as "X eats breakfast" typically happening before "X goes to work." (3) **Social Interaction (SI)**: human interactions, intentions and emotions, such as "PersonX gives a gift," leading to "PersonY feels appreciated." The complete list of relations within each group is covered in Appendix C.

To retrieve relevant commonsense knowledge,we encode the head and tail entities of all ATOMIC2020 candidates using the same BLIP2 model, then compute cosine similarities between these entity embeddings and input embeddings $f \in \{f_I, f_Q, f_C\}$. We select the top K triplets with the highest cosine similarity scores per input embedding f. This ensures only the contextually pertinent commonsense knowledge is retained, providing a solid foundation for the subsequent refinement and integration stages.

3.2 By-type Commonsense Knowledge Post-Processing

After acquiring an initial pool of commonsense triplets, we further refine them through a by-type post-processing stage, ensuring each of them is both tailored to each dataset's specific needs and contextually aligned. This stage involves two main steps: (1) By-type Commonsense Knowledge Filtering, and (2) Relevance Level Assignment.

By-type Commonsense Knowledge Filtering customizes the selection of retrieved triplets by matching the desired commonsense type distribution for each dataset. As discovered in Figure 3, each dataset benefits from a distinct mix of commonsense types. We first discard triplets with similarity scores below a threshold τ. Let $T = \{\text{CS-PE}, \text{CS-EC}, \text{CS-SI}\}$ represent the commonsense types, with each type t allocated a target proportion p_t. We then select $k_t = \lfloor p_t \times k \rfloor$ triplets from each type t with the highest similarity scores, ensuring the final set reflects the dataset's recommended distribution of commonsense knowledge. Details on these ratios are in Section 4.2.

Relevance Level Assignment further assign a qualitative relevance level to each filtered triplet based on its cosine similarity score with the input

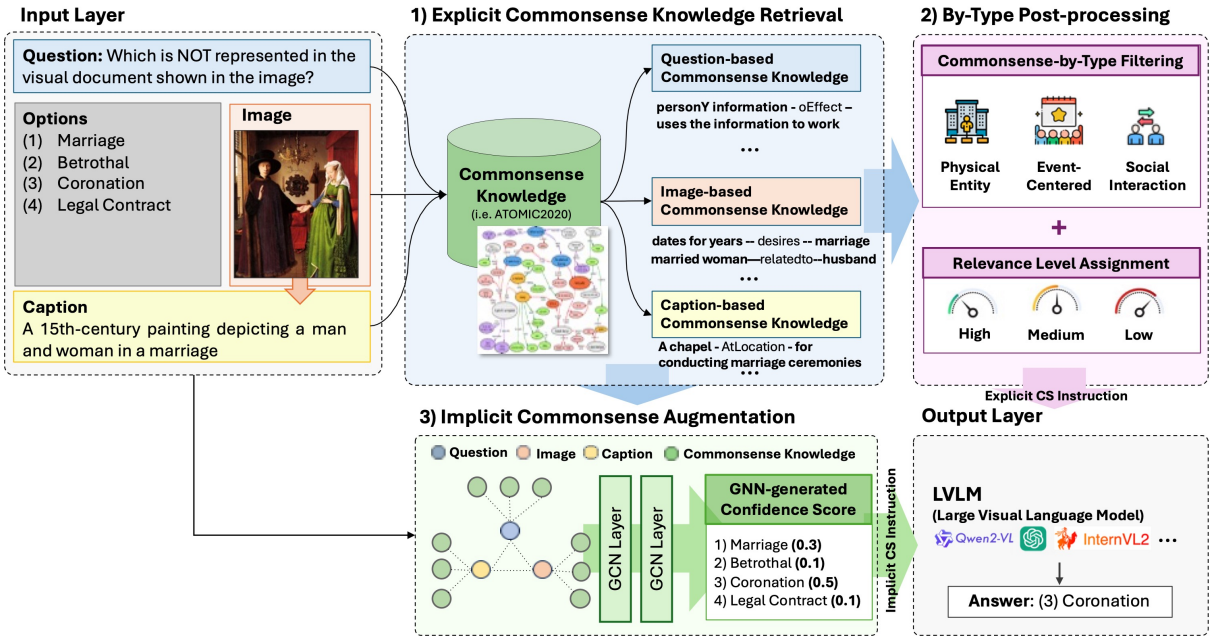


Figure 1: The proposed MAGIC-VQA Framework Architecture, which includes diverse approaches to integrate commonsense knowledge to Visual Question Answering. The detailed description of each step - 1) Explicit Commonsense Knowledge Retrieval, 2) By-Type Post-processing, 3) Implicit Commonsense Augmentation - is aligned with the subsection titles under Section 3.

sample, assisting the model in prioritizing most meaningful knowledge during reasoning. For each input source $f \in \{f_I, f_Q, f_C\}$, we first aggregate all cosine similarity scores $S_f = \{s_j^{(f)}\}$ of the selected triplets. We compute the mean μ_f and standard deviation σ_f of these scores for each dataset:

$$\mu_f = \frac{1}{N_f} \sum_{j=1}^{N_f} s_j^{(f)} \tag{1}$$

$$\sigma_f = \sqrt{\frac{1}{N_f} \sum_{j=1}^{N_f} \left(s_j^{(f)} - \mu_f \right)^2} \tag{2}$$

where N_f represents the total number of selected triplets for that input source f. As the scores have a roughly normal distribution, we apply dynamic thresholding that uses mean μ_f and standard deviation σ_f to assign each triplet a relevance level:

$$L(s_j^{(f)}) = \begin{cases} \textit{High} & \text{if } s_j^{(f)} \geq \mu_f + \frac{\sigma_f}{2} \\ \textit{Medium} & \text{if } \mu_f - \frac{\sigma_f}{2} \leq s_j^{(f)}, \\ & \text{and } s_j^{(f)} < \mu_f + \frac{\sigma_f}{2} \\ \textit{Low} & \text{if } s_j^{(f)} < \mu_f - \frac{\sigma_f}{2} \end{cases} \tag{3}$$

Detailed distribution of the similarity score for each dataset is provided in Appendix 12.

3.3 Implicit Commonsense Knowledge Augmentation

While explicit retrieval yields relevant commonsense triplets, an implicit augmentation step allows these triplets to be more deeply integrated into the reasoning process. We construct a heterogeneous graph $G_n = \{V, E\}$ where each input node (image I, question Q, and caption C) is interconnected and also linked to k additional commonsense nodes. These commonsense nodes are derived by flattening filtered commonsense triplets from Section 3.2, thereby converting each triplet into a short natural-language sentence for more straightforward integration[2]. Edges between nodes are constructed based on cosine similarity scores between their embeddings, highlighting the semantic relevance between each pair of nodes. The graph is then processed using a two-layer Graph Convolutional Network (GCN) to iteratively update node embeddings:

$$H^{(l+1)} = \rho\left(\widetilde{A} H^{(l)} W_l\right) \tag{4}$$

where ρ is a nonlinear activation function and $\widetilde{A}$ is the normalized adjacency matrix. The node embeddings $H^{(2)}$ are pooled to form a unified graph representation for each sample, which is then passed

[2]We apply a rule-based triplets flatten mechanism covered in Appendix C

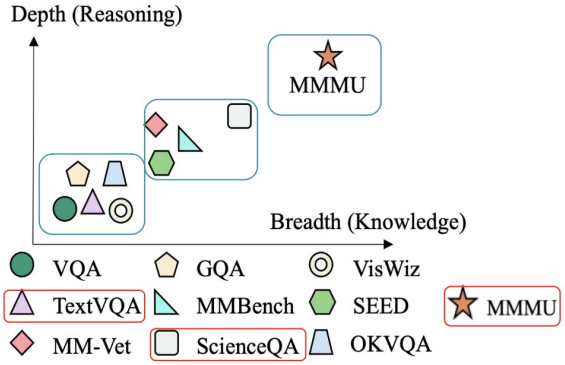


Figure 2: The comparison among VQA datasets. We selected one dataset from each of three groups. We modified the figure from (Yue et al., 2024).

through a Multi-Layer Perceptron (MLP) to produce a confidence score over candidate answers. These confidence scores provide a commonsense-augmented signal to the LVLM, enabling it to prioritize answers grounded in relevant knowledge and improving inference reliability.

3.4 Commonsense Grounded Inference

In the final inference stage, we combine all processed elements—original inputs (I, Q, C), refined commonsense triplets (with assigned relevance levels) and GNN-generated confidence scores—into a unified input structure for inference with LVLMs[3]. By fusing explicit and implicit commonsense knowledge with visual and textual signals, the LVLMs can reason effectively about nuanced relationships, delivering answers better aligned with real-world understanding.

4 Experiment

4.1 Dataset

We evaluated MAGIC-VQA on three representative VQA benchmarks of diverse complexity and depth as highlighted in Figure 2.

ScienceQA (Lu et al., 2022) comprises over 21,000 multiple-choice questions from elementary and middle school curricula in natural, social, and language science. It tests factual and procedural understanding, requiring integration of commonsense about the physical world and scientific phenomena. We select only samples with image contexts.

TextVQA (Singh et al., 2019) contains over 45,000 questions grounded in 28,000 real-world images with embedded text like signs and labels. It demands OCR to extract textual elements and integrate them with everyday commonsense provided in the context to interpret them within the visual scene. We use its validation set for our evaluation.

MMMU (Yue et al., 2024) consists of 11,550 college-level questions spanning diverse disciplines. It features challenging image types such as medical diagnosis, music sheets and so on, which goes beyond the everyday commonsense understanding emphasized in ScienceQA and TextVQA. We choose its validation set to evaluate our model.

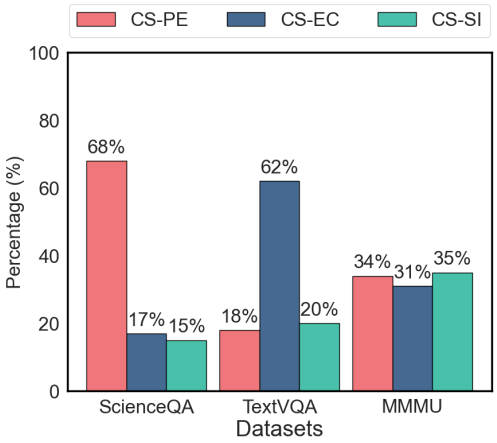


Figure 3: The distribution of categories of commonsense knowledge. CS-PE refers to physical entity-related commonsense, CS-EC to event-centered related commonsense, and CS-SI to social interaction-related commonsense.

4.2 Commonsense Knowledge Distribution

To tailor the commonsense knowledge to each dataset's specific reasoning requirements, we analyze the distribution of commonsense types across each dataset using GPT4 (OpenAI, 2024a)[4]. As Figure 3 suggests, ScienceQA requires more Physical Entity (CS-PE) knowledge, possibly due to its focus on factual and procedural scientific concepts. Meanwhile, TextVQA, which often involves contextual understanding in images, benefits more from Event-Centered (CS-EC) knowledge. MMMU, however, requires a balanced mix of Physical Entity, Event-Centered, and Social Interaction (CS-SI) commonsense due to its multi-disciplinary nature. As a result, we set the by-type filtering ratio of {CS-PE:CS-EC:CS-SI} mentioned in Section 3.2 as {0.7:0.15:0.15} in ScienceQA, {0.2:0.6:0.2} in TextVQA,{0.33:0.33:0.33} in MMMU dataset.

[3]A complete input example is provided in Appendix G

[4]The prompt template is in Appendix H

Table 2: Performance comparison under four configurations: (1) *None*: Inputs with no additional commonsense knowledge. (2) *CS Sources*: Inputs enriched with commonsense knowledge from different sources. including question (CS-Q), image (CS-I), and image caption (CS-C). (3) *CS Categories*: Inputs enriched with commonsense knowledge from different categories, including Physical Entities (CS-PE), Event-Centered (CS-EC), and Social Interaction (CS-SI). (4) *All CS*: Inputs enriched with all source-based and category-based commonsense.

Models	None	CS Sources			CS Categories			All CS
		CS-Q	CS-I	CS-C	CS-PE	CS-EC	CS-SI	
ScienceQA$_{IMG}$								
LLaVA1.6	67.50	68.83	71.56	70.35	71.12	69.01	70.83	**72.30**
BLIP3	70.00	71.56	73.88	72.97	74.03	71.57	71.05	**74.30**
InternVL2	71.99	72.58	74.37	73.91	74.56	73.09	73.21	**74.62**
Qwen2VL	71.39	72.21	74.83	71.86	74.22	72.03	72.57	**75.95**
GPT4o-mini	76.45	77.34	79.83	77.17	79.63	77.52	78.87	**81.22**
TextVQA$_{val}$								
LLaVA1.6	62.30	63.55	64.82	64.23	64.77	65.05	64.89	**65.20**
BLIP3	67.80	68.49	69.64	68.29	69.12	69.64	69.24	**69.80**
InternVL2	73.21	74.06	75.19	74.81	74.60	75.01	74.82	**75.30**
Qwen2VL	75.30	76.07	77.63	77.05	76.57	78.02	76.85	**78.90**
GPT4o-mini	78.98	79.34	81.25	80.63	80.93	81.51	81.22	**82.13**
MMMU$_{val}$								
LLaVA1.6	48.38	49.27	53.52	49.85	52.03	52.57	53.10	**54.30**
BLIP3	41.31	42.54	45.89	42.19	44.12	46.03	45.89	**47.60**
InternVL2	51.00	52.17	55.48	54.21	54.23	52.67	53.50	**55.80**
Qwen2VL	51.10	52.69	55.89	54.83	53.60	54.57	54.10	**57.42**
GPT4o-mini	55.89	56.53	58.79	56.21	58.12	57.57	57.89	**60.87**

4.3 Baselines, Metric, and Implementations

The selected baselines are four open source state-of-the-art LVLMs: LLaVA-1.6 (Liu et al., 2024a), XGen-MM (BLIP-3) (Xue et al., 2024), InternVL2 (Chen et al., 2024b), Qwen2VL (Wang et al., 2024b), and one proprietary model, GPT4o-mini (OpenAI, 2024b). These LVLMs are selected for their outstanding zero-shot performance in VQA tasks. Details of each baseline model are in Appendix D. We adopt accuracy as the evaluation metric following prior works (Singh et al., 2019; Lu et al., 2022; Yue et al., 2024). All experiments are conducted with and without the proposed MAGIC-VQA under a zero-shot setup. More implementation details are in Appendix E.

5 Results

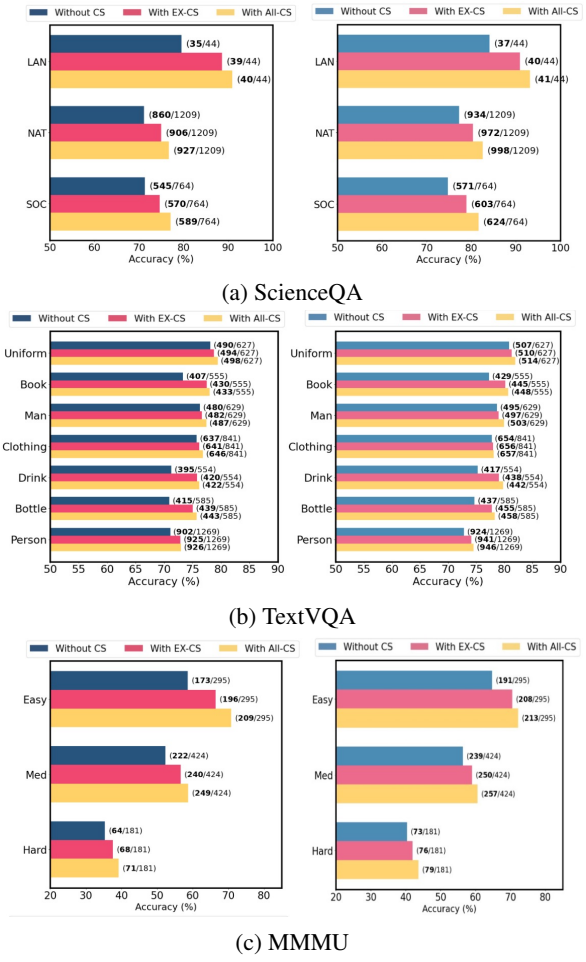
5.1 Explicit Commonsense Knowledge

We evaluated the explicit integration of commonsense knowledge triplets by systematically testing four configurations of: (1) *None*: Inputs with no additional commonsense; (2) *CS Sources*: Inputs augmented with commonsense from questions, images, or captions; (3) *CS Categories*: Inputs augmented with commonsense grouped by category (Physical Entities, Event-Centered, Social Interaction); and (4) *All CS*: Inputs augmented with all retrieved commonsense [5]. As in Table 2, integrating explicit commonsense consistently improves performance across all baselines and three datasets. For instance, on ScienceQA, GPT-4O's accuracy rises from 76.45% (*None*) to 81.22% (*All CS*), and Qwen2VL improves from 51.10% to 57.42% on MMMU under the same setup. Examining the effect of source-based commonsense reveals that image-driven knowledge (CS-I) typically provides the largest gains. For example, LLaVA1.6 on MMMU jumps from 48.38% to 53.52% with CS-I, surpassing the minor improvements from CS-Q or CS-C. This suggests that leveraging image-aligned commonsense offers more grounded cues for inference. However, category-based commonsense (CS-PE, CS-EC, and CS-SI) exhibits dataset-dependent effectiveness. On ScienceQA, CS-PE is most beneficial, while on TextVQA, CS-EC dominates, and MMMU shows a more balanced pattern. These results align with our earlier commonsense distribution analysis in Section 4.2, highlighting the importance of tailoring knowledge retrieval to the dataset's unique characteristics.

[5] Each experiment is tested with a fixed number of $k = 6$ to maintain a fair comparison.

Model	Ex-CS		Im-CS	Performance		
	CS	Rel	Conf	SQA	MMMU	TVQA
Qwen2VL	✗	✗	✗	71.39	51.10	75.30
Qwen2VL	✓	✗	✗	75.11	56.00	78.50
Qwen2VL	✗	✗	✓	72.88	53.41	76.42
Qwen2VL	✓	✓	✗	75.95	57.42	78.90
Qwen2VL	✓	✗	✓	76.42	57.21	79.10
Qwen2VL	✓	✓	✓	**77.12**	**58.72**	**79.80**
GPT4o-mini	✗	✗	✗	76.45	55.89	78.98
GPT4o-mini	✓	✗	✗	80.07	59.30	81.73
GPT4o-mini	✗	✗	✓	77.02	57.64	79.55
GPT4o-mini	✓	✓	✗	81.22	60.87	82.13
GPT4o-mini	✓	✗	✓	80.94	60.25	82.50
GPT4o-mini	✓	✓	✓	**82.50**	**61.03**	**83.37**

Table 3: Quantitative analysis on the effect of each component of MAGIC-VQA on the model performance. The "Ex-CS" (CS and Rel) denotes explicit commonsense knowledge inclusion (All-CS in Section 5.1), while "Im-CS" ($Conf$) denote implicit commonsense inclusion. Green check (✓) denotes the inclusion of a component, and red cross (✗) denotes exclusion.



(a) ScienceQA

(b) TextVQA

(c) MMMU

Figure 4: Subcategory-level accuracy on (a) ScienceQA, (b) TextVQA, and (c) MMMU for Qwen2VL (left) andGPT4o-mini (right) under three conditions: Without CS, With EX-CS and With All-CS.

5.2 Implicit Commonsense Knowledge

We next examined the effect of implicit commonsense knowledge augmentation as outlined in Section 3.3 using two representative LVLMs, Qwen2VL andGPT4o-mini. We also include results from explicit commonsense knowledge integration (All-CS in Section 5.1). As Table 3 suggests, while the implicit commonsense knowledge $Conf$ does not contribute as significantly as explicit commonsense knowledge (*Ex-CS*), it nonetheless provides complementary information that enhances overall performance. Incorporating only $Conf$ with Qwen2VL improves the MMMU accuracy from 51.10% to 53.41%, and withGPT4o-mini, the accuracy increases from 55.89% to 57.64%. We also observe that adding Rel notably improves results across all three datasets. Furthermore, combining implicit and explicit commonsense yields the highest overall performance, indicating that implicit augmentation complements explicit knowledge by capturing additional nuances and context that explicit methods alone may miss. Further detailed qualitative analysis is provided in Section 6 and the Appendix B.

5.3 Break Down Results

We compare performance on specific subcategories within each selected dataset using Qwen2VL and GPT4o-mini in order to analyze the effects of commonsense knowledge augmentation in Figure 4. Across all datasets and subcategories, incorporating commonsense significantly improves the accuracy. Each dataset features distinct subcategories that would benefit from varying aspects of commonsense reasoning.

First, **ScienceQA** in Figure 4a, commonsense augmentation yields notable improvements, particularly in language-related subcategories, reflecting the value of context-sensitive reasoning. **TextVQA** in Figure 4b, categories involving concrete objects, such as *'uniform'* and *'books'*, benefit more significantly from commonsense augmentation compared to abstract categories like *'persons'*, indicating that concrete objects allow for more precise retrieval of

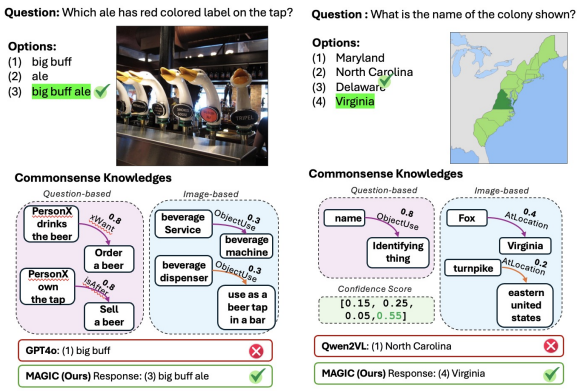
relevant commonsense knowledge. For **MMMU** in Figure 4c, commonsense augmentation benefits *easy*-level questions, closely tied to everyday knowledge, while struggles with *hard*-levels that demand complex reasoning beyond commonsense.

5.4 Further Ablation Studies

Beyond our primary experiments, we conducted additional ablation studies to further analyze the impact of various design choices and methodological components. The detailed results for these studies are available in Appendix A. Firstly, as in Appendix A.1, we examined the effect of varying the selection ratio of different commonsense knowledge types—CS-PE, CS-EC, and CS-SI—when integrated with Qwen2VL on the ScienceQA. This analysis helps to determine how the balance of different types of commonsense knowledge influences model performance. Secondly, in Appendix A.2, we investigated the impact of incorporating different numbers of commonsense knowledge triplets using both Qwen2VL and GPT4o-mini. This study aims to assess whether increasing the number of commonsense knowledge triplets enhances reasoning capabilities or if there is a saturation point beyond which performance gains plateau. Thirdly, Appendix A.3 presents the effect of diverse similarity metrics by comparing Manhattan, Cosine, and Euclidean Distance when applied with GPT4o-mini, providing insights into how different similarity measures affect retrieval effectiveness. Finally, as presented in Appendix A.4, we conducted an evaluation of VLPM-style fine-tuning by comparing our proposed approach with methods that distill implicit commonsense knowledge using graph-bassed techniques into compact VLPMs, such as ViLT and ALBEF. This comparison highlights the advantages of our method in effectively integrating commonsense reasoning within smaller VLPMs while maintaining performance efficiency.

6 Qualitative Analysis

Figure 5 compares MAGIC-VQA with GPT4o and Qwen2VL, demonstrating how our framework effectively integrates both explicit and implicit commonsense knowledge for enhanced visual question answering. As illustrated in Figure 5a, while GPT4o struggles to deduce the complete answer (big buff ale) to input query, MAGIC-VQA successfully incorporate contextual knowledge, such as "Person X owns the tap sells beer" and "beverage dispenser used as a beer tap in a bar," linking beer consumption, tap functionality, and beverage machines with the input question to arrive at the correct answer. In Figure 5b, while Qwen2VL incorrectly identifies the colony as North Carolina, our MAGIC-VQA addresses this limitation by integrating explicit image-based commonsense knowledge about Virginia's location, historical turnpikes, and wildlife, correctly concluding that the answer is Virginia. Confidence scores derived from implicit commonsense knowledge further reinforce the evidence for the final accurate prediction. Additional case studies are shown in Appendix B.



(a) TextVQA example (b) ScienceQA example

Figure 5: Comparison of results of commonsense knowledge-injected MAGIC-VQA (ours) and original GPT4o-mini and Qwen2VL across different datasets, including TextVQA and ScienceQA. Each example highlights the question-based and image-based explicit commonsense knowledge. Example in ScienceQA is also injected with implicit commonsense knowledge.

7 Conclusion

This paper introduced MAGIC-VQA, a novel framework integrating commonsense knowledge into VQA to address the limitations of existing LVLMs. MAGIC-VQA's three-stage process—knowledge retrieval, commonsense post-processing, and GNN-based augmentation—enables nuanced reasoning without extensive pre-training or complex prompt tuning. Evaluations on ScienceQA, TextVQA, and MMMU demonstrate significant improvements in tasks requiring advanced reasoning. This framework establishes a robust approach for bridging raw visual inputs with high-level reasoning, offering scalable enhancements for VQA. We hope this work inspires further research into structured commonsense reasoning for complex multimodal challenges.

Limitation

While the MAGIC-VQA framework demonstrates significant improvement, it currently relies on external knowledge graphs, such as ATOMIC2020 and predefined commonsense categories, which may limit its adaptability to diverse and unforeseen domains. Additionally, real-world VQA scenarios often involve noisy or ambiguous inputs that may not always align with the structured assumption of the commonsense knowledge graph. To address these limitations, we plan to extend our approach by developing and incorporating a more diverse and extensive range of multimodal commonsense knowledge sources. Expanding the scope of knowledge representation will enhance multimodal understanding and learning ability and help us handle more multimodal reasoning tasks.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Annie S Chen, Alec M Lessing, Andy Tang, Govind Chada, Laura Smith, Sergey Levine, and Chelsea Finn. 2024a. Commonsense reasoning for legged robot adaptation with vision-language models. *arXiv preprint arXiv:2407.02666*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on clip via vision-language knowledge distillation. *Preprint*, arXiv:2203.06386.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. 2022. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5089–5098.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *Preprint*, arXiv:1609.02907.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Zhenyang Li, Yangyang Guo, Kejie Wang, Xiaolin Chen, Liqiang Nie, and Mohan Kankanhalli. 2023b. Do vision-language transformers exhibit visual commonsense? an empirical study of vcr. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5634–5644.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18798–18806.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jae Sung Park, Jack Hessel, Khyathi Chandu, Paul Pu Liang, Ximing Lu, Peter West, Youngjae Yu, Qiuyuan Huang, Jianfeng Gao, Ali Farhadi, et al. 2024. Localized symbolic knowledge distillation for visual commonsense models. *Advances in Neural Information Processing Systems*, 36.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Sahithya Ravi, Aditya Chinchure, Leonid Sigal, Renjie Liao, and Vered Shwartz. 2023. Vlc-bert: Visual question answering with contextualized commonsense knowledge. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1155–1165.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024a. T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19162–19170.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qunbo Wang, Ruyi Ji, Tianhao Peng, Wenjun Wu, Zechao Li, and Jing Liu. 2024c. Soft knowledge prompt: Help external knowledge become a better teacher to instruct llm in knowledge-based vqa. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6132–6143.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Visually-augmented language modeling. *Preprint*, arXiv:2205.10178.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2022. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.

Shuo Yang, Siwen Luo, and Soyeon Caren Han. 2024. Multimodal commonsense knowledge distillation for visual question answering. *arXiv preprint arXiv:2411.02722*.

S Ye, Y Xie, D Chen, Y Xu, L Yuan, C Zhu, and J Liao. Improving commonsense in vision-language models via knowledge graph riddles (2022). *URL: https://arxiv. org/abs/2211.16504. doi*, 10.

Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, and Jing Liao. 2023. Improving commonsense in vision-language models via knowledge graph riddles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2634–2645.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. *Preprint*, arXiv:2205.01850.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*.

A Additional Experiment Results

A.1 Ratio of Commonsense Knowledge Type

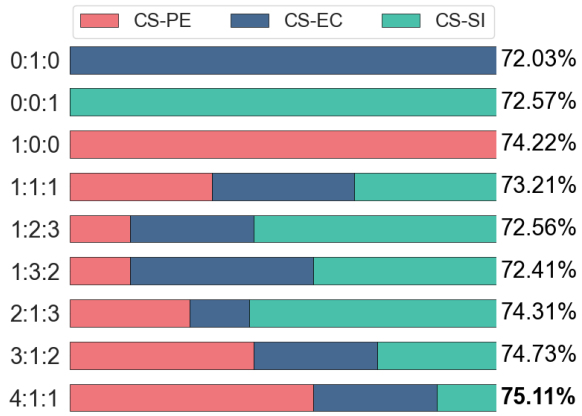
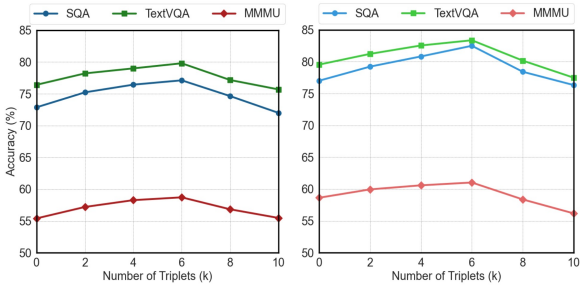


Figure 6: Effect of the ratio of commonsense knowledge categories on Qwen2VL. The three categories of explicit commonsense knowledge include social interactions (CS-SI), physical entities (CS-PE), and event-based relations (CS-EC).

To determine the optimal ratio of commonsense knowledge types, we conduct an experiment to analyze how varying distributions of CS-PE, CS-EC, and CS-SI knowledge triplets impact the performance of Qwen2-VL on the ScienceQA dataset. Figure 6 indicates that the model achieves its highest accuracy of 75.11% with a 4:1:1 ratio, which places greater emphasis on CS-PE knowledge. Additionally, distributions favoring CS-PE triplets consistently result in improved performance. For example, a 3:1:2 ratio achieves an accuracy of 74.73%. In contrast, ratios prioritizing CS-EC or CS-SI, such as 1:2:3 or 1:3:2, yield lower accuracies of 72.56% and 72.41%, respectively. These results suggest that CS-PE is the most essential commonsense knowledge type for the ScienceQA, aligning with its focus on physical concepts and entities as discussed in Section 5.1. Notably, the 4:1:1 ratio closely mirrors the inherent distribution of commonsense knowledge in ScienceQA in Figure 3. This alignment suggests that tailoring the balance of commonsense knowledge to match the dataset's inherent characteristics, as demonstrated by our approach, leads to the most significant performance improvements.

A.2 Number of Knowledge Triplets

To optimize the integration of explicit commonsense knowledge into our model while minimizing the risk of introducing excessive noise, we investigate how varying the number of retrieved commonsense triplets k impacts performance. We vary k from 0 to 10, incrementally increasing the number of triplets provided to Qwen2VL andGPT4o-mini and measuring the corresponding accuracy. In Figure 7, the models achieve optimal performance when $k = 6$ triplets are incorporated. Utilizing fewer than $k = 6$ appears insufficient to provide the contextual information for efficient reasoning, while exceeding $k = 6$ triplets includes irrelevant or redundant information, diminishing performance. Therefore, using $k = 6$ represents the optimal balance/sweet spot for enriching the model with necessary knowledge while focusing on pertinent information.



(a) Qwen2VL (b) GPT4o-mini

Figure 7: Effect of number of knowledge triplets k: Comparison between Qwen2VL andGPT4o-mini.

A.3 Effect of Similarity Metric

We further conduct an ablation study to evaluate the effect of different retrieval metrics on explicit commonsense knowledge retrieval across different datasets and input sources using GPT4-O model. We explore the performance of three retrieval metrics: *Cosine Distance*, *Manhattan Distance*, and *Euclidean Distance* to understand their influence on retrieval efficacy across these datasets and input types.

As suggested on the right Figure 8, the performance trends reveal notable differences in retrieval effectiveness depending on the metric and dataset. For ScienceQA, *Manhattan Distance* achieves the highest scores. Meanwhile, the MMMU dataset shows relatively low and uniform scores across all metrics, suggesting that this dataset's retrieval performance is less sensitive to the choice of metric, potentially due to the diversity and complexity of MMMU's multimodal inputs.

When comparing retrieval metrics across different input sources, we observe further variations.

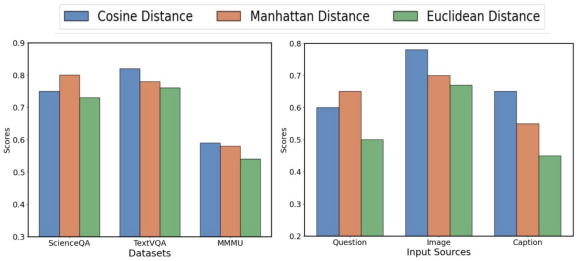
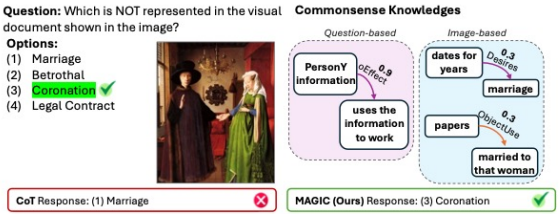


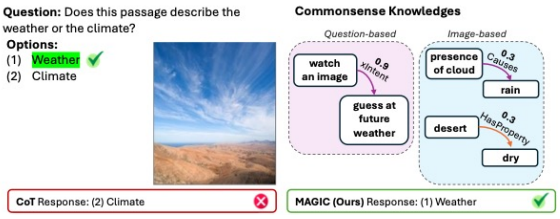
Figure 8: Comparison of accuracy across three datasets (ScienceQA, TextVQA, MMMU) using different distance metrics (Cosine, Manhattan, Euclidean) for explicit commonsense retrieval (Left). The right panel illustrates the performance of these metrics across different input sources (Question, Image, Caption) within the ScienceQA dataset.

Model	I	Q	C	SQA	MMMU	TextVQA
ViLT	-	-	-	56.14	23.04	41.49
MAGIC-VQA$_{(ViLT)}$	✗	✓	✓	60.53	20.12	40.13
MAGIC-VQA$_{(ViLT)}$	✓	✗	✓	53.32	14.28	32.24
MAGIC-VQA$_{(ViLT)}$	✓	✓	✗	63.45	22.37	43.98
MAGIC-VQA$_{(ViLT)}$	✓	✓	✓	**65.41**	**23.35**	**44.12**
ALBEF	-	-	-	59.12	25.38	39.27
MAGIC-VQA$_{(ALBEF)}$	✗	✓	✓	61.24	24.43	37.28
MAGIC-VQA$_{(ALBEF)}$	✓	✗	✓	57.42	17.21	28.42
MAGIC-VQA$_{(ALBEF)}$	✓	✓	✗	66.79	26.91	42.88
MAGIC-VQA$_{(ALBEF)}$	✓	✓	✓	**68.33**	**27.32**	**43.25**

Table 4: Combined Results with Multimodal Contributions. The green checkmarks (✓) denote the inclusion of a component, while the red crosses (✗) denote its exclusion.



(a) The case from MMMU.



(b) The case from ScienceQA.

Figure 9: Visualisations of MAGIC-VQA results on the MMMU and ScienceQA datasets, showcasing the role of image-based commonsense knowledge in deriving correct answers. This highlights the cases when image-based commonsense knowledge is more influential in finding the answer.

For question-based retrieval, *Manhattan Distance* consistently yields higher performance scores, indicating that the absolute differences in feature spaces may be more informative for question-centric retrieval. In contrast, image-based and caption-based retrieval achieves the highest scores with *Cosine Distance*, suggesting that angle-based similarity is more effective for capturing visual context in the knowledge graph.

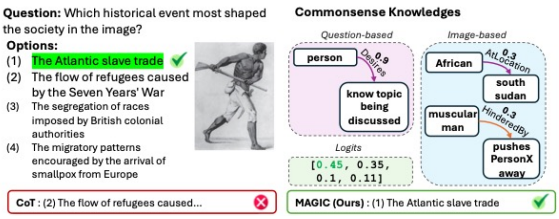
A.4 VLPM-Style Fine-Tuning Results

We further perform VLPM-style fine-tuning to investigate the effectiveness of implicit multimodal commonsense in MAGIC-VQA framework. Table 4 presents a quantitative analysis of the impact of different input nodes—image (I), question (Q), and generated caption text (C)—on the performance of two baseline models (VILT and ALBEF) in VLPM-style fine-tuning. Each node type corresponds to a specific input modality, and when removed, the original embeddings from the pre-trained models are used instead of GCN-trained node embeddings. It is found that including all nodes consistently yields the best outcomes, underscoring the complementary contributions of visual, textual and caption inputs multimodal reasoning.

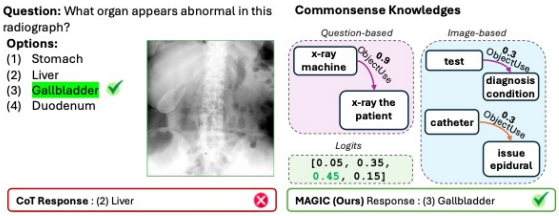
Notably, the question node proves to be the most crucial across all nodes. For example, ALBEF's accuracy on ScienceQA declines from 68.33% to 57.42% when the question node is excluded, highlighting its essential role in guiding the model's attention toward relevant aspects of the image and improving reasoning. On the other hand, the image node also plays a significant role in performance. Visual inputs provide critical scene-level information, enabling models to capture object attributes and spatial relationships, which cannot be fully compensated by text-based inputs alone.

B Additional Case Studies

We summarize different qualitative analysis case studies in three types: **1)** The cases when the image-based explicit commonsense knowledge plays an essential role (Figure 9), **2)** The cases when implicit commonsense-based confidence plays an essential role (Figure 10), and **3)** the cases when by-type commonsense knowledge post-processing plays an important role (Figure 11).

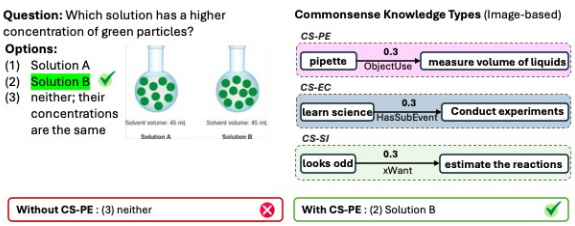


(a) The case from MMMU.

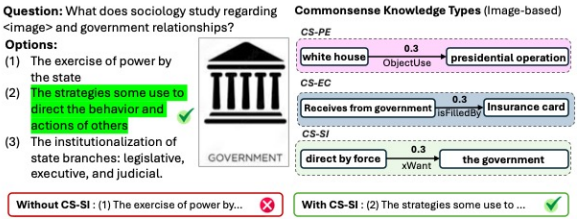


(b) The case from MMMU.

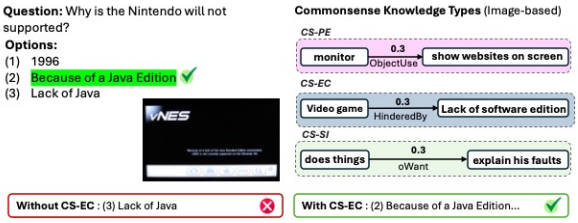
Figure 10: Visualisation of MAGIC-VQA results on MMMU datasets. This highlights the cases when implicit commonsense-based confidence plays an essential role.



(a) The ScienceQA case when CS-PE is influential



(b) The ScienceQA case when CS-SI is influential



(c) The TextVQA case when CS-PE is influential

Figure 11: Visualisation of MAGIC-VQA results on ScienceQA and TextVQA datasets. This highlights the cases when by-type commonsense knowledge post-processing plays an important role

C Commonsense Relation Transformation Table

We use *"Someone"*, *"Someone's"* to replace the *"PersonX"*, *"PersonX's"* and *"Another"*, *"Another one's"* to replace *"PersonY"*, *"PersonY's"* separately, in the heads and tails of the Atomic2020 triplets to enhance clarity and coherence in commonsense-grounded inference.

Relation	Transformed Format
	Physical-Entity
ObjectUse	is used for
AtLocation	is at
MadeUpOf	is made up of
HasProperty	can be
CapableOf	is capable of
Desires	desires
NotDesires	does not desire
	Event-Centered
IsAfter	occurs after
HasSubEvent	has sub-event
IsBefore	occurs before
HinderedBy	is hindered by
Causes	causes
xReason	is because someone
isFilledBy	is filled by
	Social-Interaction
xNeed	then someone needs
xAttr	then someone has attributes
xEffect	then someone has the effect
xReact	then someone reacts with
xWant	then someone wants
xIntent	then someone intends
oEffect	then the effect on another is
oReact	then another reacts with
oWant	then another one wants

Table 5: Transformation template of commonsense relations into natural language phrases

D Baselines

- **LLaVA-1.6 (LLaVA-Next)** (Li et al., 2024): is an open-source Large Multimodal Model (LMM) designed for enhanced visual and conversational understanding built upon LLaVA (Liu et al., 2024b). It supports higher input resolutions (up to 672x672 pixels) for finer visual detail recognition and incorporates improved visual instruction tuning for better reasoning and OCR capabilities. LLaVA 1.6 is highly efficient, using fewer than 1 million visual instruction tuning samples and a streamlined training process. Its versatility enables it to handle a wide range of applications, from image and text-based tasks to complex multimodal interactions, all while maintaining a minimalist and data-efficient design. We use llava-v1.6-mistral-7b-hf checkpoints for zero-shot testing.

- **InternVL2** (Chen et al., 2024b): is a state-

of-the-art multimodal large model developed by OpenGVLab. It integrates image, video, text, speech, and 3D data, supporting over 100 tasks with exceptional performance across benchmarks. InternVL2 leverages a progressive alignment training strategy and have achieved outstanding results in complex multimodal understanding tasks, rivaling leading commercial closed-source models like GPT-4V (OpenAI, 2024a). It introduces innovations like vector linking for diverse outputs. It also has parameter sizes ranging from 1B to 76B optimized for efficiency, which delivers high performance even on limited resources. We use InternVL2-8B-hf checkpoints for zero-shot testing.

- **xGen-MM(BLIP-3)** (Xue et al., 2024): is a cutting-edge framework for Large Multimodal Models (LMMs) developed by Salesforce AI Research. It features a modular architecture with a scalable vision token sampler and a pre-trained language model, optimized for diverse multimodal tasks such as image captioning, visual question answering, and OCR. It simplifies training objectives with a unified auto-regressive loss and incorporates post-training techniques like Direct Preference Optimization (DPO) and safety fine-tuning to improve truthfulness and mitigate harmful behaviors. We use xgen-mm-phi3-mini-instruct-r-v1 checkpoints for zero-shot testing.

- **Qwen2-VL** (Wang et al., 2024b): is a cutting-edge vision-language model designed with a robust technical architecture to process multimodal inputs efficiently. It integrates a 675M-parameter Vision Transformer (ViT) enhanced with a Naive Dynamic Resolution mechanism, enabling adaptive encoding of images and videos into variable-length visual tokens to capture detail at multiple scales. To align spatial and temporal information, the model employs Multimodal Rotary Position Embedding (M-RoPE), decomposing positional information into temporal, height, and width dimensions. It also leverages dynamic sequence lengths and efficient parallelism techniques, allowing for deployment in sizes of 2B, 7B, and 72B parameters. We use Qwen2-VL-7B-Instruct checkpoints for zero-shot testing.

- **GPT-4o** (OpenAI, 2024b): is an advanced autoregressive model that processes and generates multimodal content, including text, images, audio, and video, using a unified neural network architecture. It offers significant enhancements in vision and audio understanding, multilingual text generation, and operational efficiency. The model's training incorporates diverse public and proprietary datasets across modalities, with rigorous post-training alignment to ensure safety and mitigate risks such as bias, misinformation, and unauthorized content generation. We use gpt-4o-2024-08-06 checkpoints for zero-shot testing.

E Implementation Details

We set $K = 30$ to retrieve explicit commonsense knowledge. For by-type commonsense knowledge processing, we configure $\varepsilon = 0.1$ and $k = 6$ to effectively integrate rich commonsense knowledge while minimizing the introduction of excessive noise. For implicit commonsense confidence augmentation, we follow the default setup in Kipf and Welling (2017) to explore a standard two-layer GCN. The dimension of the hidden size is set to be 256 and 512, each followed by a dropout layer with the rate to be 0.4. To train the teacher model, we explore batch size to be 64, learning rate to be 1e-5 and epoch to be 30 with early stopping for all models and datasets. There is a global average pooling layer and a output layer using the softmax function after the last GCN layer.

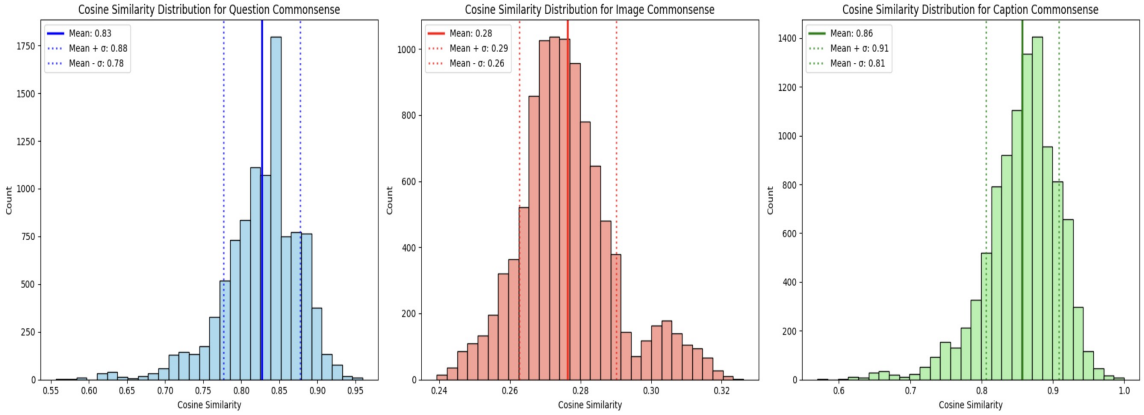
All experiments are conducted on a workstation equipped with one A100 GPU with 40 GB of VRAM. We utilize PyTorch 1.10.0 for model training and the HuggingFace Transformers library for accessing pre-trained models. Our code was written in Python 3.8, and CUDA 11.2 was used for GPU acceleration.

F Retrieved Commonsense Triplet Cosine Similarity Distribution

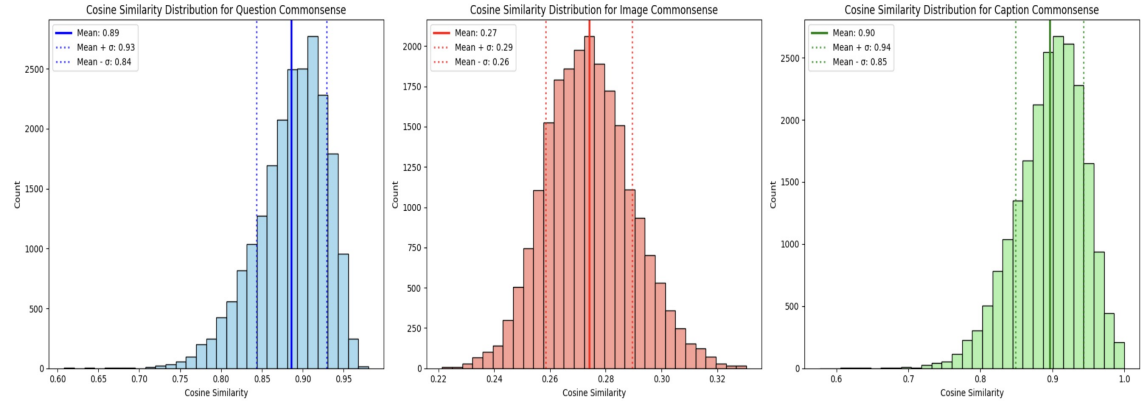
Figure 12 depicts the cosine similarity distribution of the retrieved triplets of each input source across all three datasets.

G Concrete Example of Input Prompt

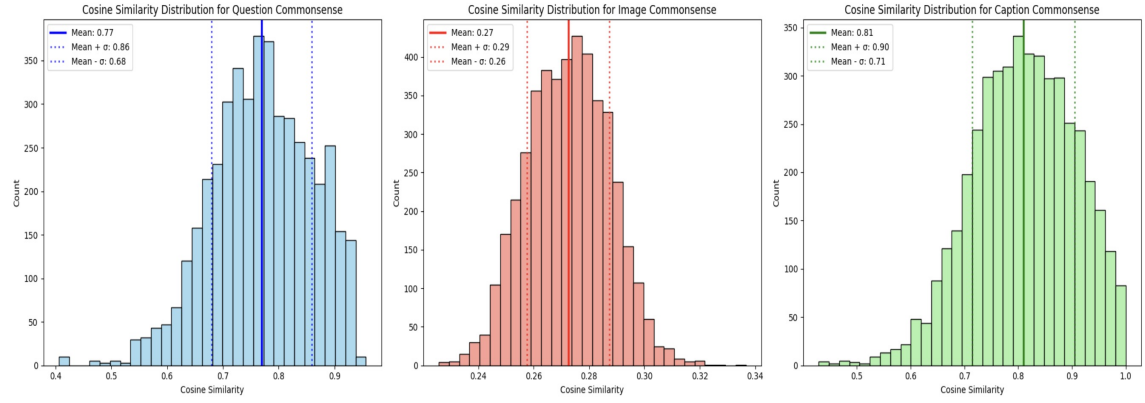
We further include a concrete prompt example of our MAGIC-VQA framework. as demonstrated in in Table 6, the input sample is augmented with both explicit and implicit commonsense knowledge providing the background information. We ask the



(a) ScienceQA



(b) TextVQA



(c) MMMU

Figure 12: Overall cosine similarity distributions for three input sources within each dataset. The first column represents the cosine similarity distribution of retrieved triplets for input question. The second column represents the cosine similarity distribution of retrieved triplets for input image. The third column represents the cosine similarity distribution of retrieved triplets for input caption.

model to first generate the rational then answer the question.

H Commonsense Category Analysis Prompt Format

To analyze the commonsense knowledge distribution within each selected dataset, we provide the following prompt template to classify each sample to their most relevant commonsense knowledge (CS-PE, CS-EC, CS-SI) covered in Table 7.

Background

You are an advanced Vision-Language Model assistant designed to answer multiple-choice questions based on a given image. Your task is to select the most appropriate option from the provided answer choices. You are given an input image, a question related to the image, the image caption, multiple-choice answer options, and both explicit and implicit commonsense knowledge.

Explicit commonsense knowledge consists of statements related to the input, categorized as image-related commonsense, question-related commonsense, and caption-related commonsense. Implicit commonsense knowledge includes the relevance level (e.g., highly relevant, relevant, less relevant) assigned to each explicit commonsense statement and the confidence of each candidate option, where higher values indicate a greater likelihood of being correct.

Your objective is to integrate the explicit and implicit commonsense knowledge with the provided information to generate a step-by-step reasoning. Based on this rationale, you will select the most appropriate answer from the given options.

Input Information

Image:

Question: Which of these oceans does the prime meridian intersect?
Caption: An image of a world map with labeled continents and oceans.
Options:
A. "the Atlantic Ocean"
B. "the Indian Ocean"
C. "the Pacific Ocean"

Explicit Commonsense Knowledge

Image-Related Commonsense:
- The Atlantic Ocean is at the western hemisphere. (Highly Relevant)
- A world traveler is capable of crossing many time zones. (Relevant)

Question-Related Commonsense:
- A traveler is capable of crossing geographical borders. (Highly Relevant)
- Someone who is far from home might want to measure the distance. (Less Relevant)

Caption-Related Commonsense:
- The Atlantic Ocean is used for separating continents. (Highly Relevant)
- If someone sees the ocean, they might think of traveling to it. (Relevant)

Implicit Commonsense Knowledge (Confidence for Each Option)
A: 0.6
B: 0.05
C: 0.35

Rationale:

Answer:

Table 6: A concrete example of the input prompt

Prompt Template for Commonsense Category Classification

Instructions:
You are an expert in commonsense reasoning and knowledge representation. Your task is to classify each sample into one of three commonsense categories:
1. **Physical-Entity Commonsense (CS-PE)**: Knowledge about physical objects, their properties, uses, locations, and physical attributes. This includes understanding what things are made of, typical or atypical uses, and physical characteristics.
2. **Event-Centered Commonsense (CS-EC)**: Knowledge about events, including their causes, effects, prerequisites, sequences, and hindrances. This encompasses understanding how events are related in time and causality.
3. **Social-Interaction Commonsense (CS-SI)**: Knowledge about social behaviors, mental states, interactions, and interpersonal dynamics. This involves understanding intentions, emotional reactions, and attributes in social contexts.
Sample:

- **Image:** $< ImageCaption >$
- **Question:** $< Question >$
- **Choices:** $< Options >$
- **Answer:** $< Answer >$

Reasoning Steps:
Please first examine the question and answer choices, along with the image caption, to identify the main focus of the sample. Then provide a step-by-step reasoning on how specific elements of the sample align with the potential commonsense category. Then assign the appropriate commonsense category (CS-PE, CS-EC, or CS-SI) based on the provided rationale.
Classification:

Table 7: Prompt Template for Classifying Samples into Commonsense Categories