

KRAKEN: Knowledge Representation with Augmented Knowledge Graph Encoding for Neural VQA

Shubham Goel
IIIT Hyderabad, Hyderabad
Telangana, India

Nidhi Vaidya
IIIT Hyderabad, Hyderabad
Telangana, India

Shree Mitra
IIIT Hyderabad, Hyderabad
Telangana, India

Abstract

Visual Question Answering (VQA) models excel at perceptual tasks but struggle with questions that require external, structured knowledge. While Knowledge Graph Question Answering (KGQA) has emerged to address reasoning over symbolic data, existing methods are ill-equipped to handle multimodal inputs that blend visual and textual information with structured graphs. To bridge this gap, we introduce a novel multimodal reasoning architecture. Our model is designed to reason at the intersection of language, structured knowledge, and visual information. It integrates three specialized pathways: (1) a frozen Vision Transformer (ViT) coupled with a trainable Q-Former to extract salient visual features, (2) a dedicated graph encoder to produce a topologically-aware representation of relational facts, and (3) an input text encoder. The outputs of these pathways are projected into a common space and fed to a frozen Large Language Model (LLM), which acts as the central reasoning engine. We employ a two-stage training strategy, first aligning the modalities and then fine-tuning the connective components with Parameter-Efficient Fine-Tuning (PEFT). Our work presents not just a new model for graph-based QA, but a novel architecture for the more challenging task of multimodal, knowledge-augmented reasoning.

1 Introduction

Visual Question Answering (VQA) has made significant strides in enabling machines to answer questions about the content of an image. Modern systems can successfully identify objects, describe attributes, and count instances, demonstrating a strong capacity for visual perception. However, a critical frontier remains: answering questions that require knowledge beyond the pixels of the image itself. For instance, while a model might identify a picture of the Eiffel Tower, it cannot answer "Who designed the landmark in this photo?" without access to external, structured world knowledge.

To address this need for factual reasoning, the field of Question Answering over Knowledge Graphs (KGQA) has developed sophisticated methods to query large-scale KGs (Saxena et al., 2020; Schlichtkrull et al., 2017). The paradigm has evolved from learning embeddings and parsing questions into formal queries (Gu et al.) to leveraging the powerful in-context reasoning and tool-use capabilities of Large Language Models (LLMs) (Wei et al., 2023; Cheng et al., 2023). Despite their success, these KGQA systems share a fundamental limitation: they operate exclusively on textual questions and symbolic graph structures, lacking a pathway to incorporate visual information. This leaves a crucial gap at the intersection of vision, language, and structured knowledge.

In this paper, we address this gap by proposing a novel multimodal architecture designed explicitly for knowledge-augmented reasoning. Our model is built on the insight that true comprehensive understanding requires the intelligent fusion of heterogeneous data sources. We situate our work within the modern LLM-based paradigm, using a powerful, pre-trained LLM as our central reasoning engine. However, we introduce a critical and novel extension: a multimodal input system that equips the LLM with a richer, more grounded context. Our architecture integrates three distinct pathways: a frozen Vision Transformer (ViT) from CLIP (Radford et al., 2021) for visual understanding, a dedicated graph encoder for symbolic relational facts, and a pathway for the user’s textual question. To bridge the vision and language modalities effectively, we employ a Q-Former inspired by the BLIP-2 architecture (Li et al., 2023), which distills the visual information into a compact set of salient features.

Our contributions are threefold:

1. We introduce a novel end-to-end multimodal architecture that, for the first time, fuses sym-

bolic graph representations, visual features, and text for knowledge-based VQA.

2. We detail a stable and efficient two-stage training strategy involving a modality alignment phase followed by Parameter-Efficient Fine-Tuning (PEFT) of the connective components.
3. We present an empirical analysis demonstrating the limitations of a vision-only approach to understanding rendered knowledge graphs, thereby motivating the necessity of our multimodal design.

2 Related Works

Research in Question Answering over Knowledge Graphs (KGQA) has evolved through several distinct paradigms, each aiming to more effectively bridge the gap between natural language questions and structured knowledge. Our work builds upon insights from these successive approaches, culminating in a multimodal architecture that leverages the reasoning power of modern Large Language Models (LLMs).

Early deep learning approaches focused on learning dense vector representations of both the question and the knowledge graph components. A prominent example of this is **EmbedKGQA** (Saxena et al., 2020), which frames the task as a retrieval problem within a learned embedding space. By training a model to score the proximity of question embeddings to candidate entity embeddings, this method can effectively answer complex multi-hop questions. Foundational to this line of work are powerful graph representation learning techniques like the **Relational Graph Convolutional Network (RGCN)** (Schlichtkrull et al., 2017). RGCNs extend traditional GNNs to handle the heterogeneous, multi-relational nature of KGs, producing rich node and edge embeddings that capture the graph’s topology. While powerful, these embedding-based methods often perform reasoning implicitly and can struggle with generalization to unseen entities or relations.

A parallel and highly effective paradigm is **Semantic Parsing**, which seeks to translate a natural language question into a formal, executable query. Instead of learning embeddings, these models learn a direct mapping to a language like SPARQL. The **GrailQA** (Gu et al.) benchmark and its associated models demonstrated the power of this approach, showing remarkable zero-shot generaliza-

tion to KG schemas and compositional structures unseen during training. By generating an explicit, interpretable query, these methods offer high precision but can be sensitive to the complexity and variability of natural language.

The recent advent of Large Language Models (LLMs) has fundamentally shifted the landscape of KGQA. The immense world knowledge and emergent reasoning capabilities of models like GPT-4 have enabled new, highly flexible approaches. One strategy is to use LLMs directly as zero-shot reasoners through sophisticated prompting, such as **Chain-of-Thought (CoT)** (Wei et al., 2023). CoT prompting encourages the model to break down a question into intermediate logical steps, significantly improving its performance on complex reasoning tasks without any task-specific fine-tuning. A more advanced strategy involves fine-tuning LLMs to function as intelligent agents that can use external tools. Models like **Binder** (Cheng et al., 2023) and **StructGPT** (Jiang et al., 2023b) are trained to generate and execute code or API calls to query structured data sources, including KGs. This "tool use" paradigm allows the LLM to offload factual retrieval to the KG while focusing its own capacity on reasoning, planning, and synthesizing the retrieved information.

Our proposed work situates itself within this latest LLM-based paradigm, using a powerful LLM as our central reasoning engine. However, we introduce a critical and novel extension: **multimodality**. Whereas the aforementioned methods operate exclusively on textual questions and symbolic graph structures, our model is designed to reason at the intersection of language, structured knowledge, and visual information. By integrating a dedicated graph encoder for relational facts and a sophisticated vision pathway (ViT + Q-Former) for image understanding, we equip the LLM with a richer, more grounded context. Our contribution is therefore not just a new model for graph-based QA, but a novel architecture for a more challenging task: multimodal, knowledge-augmented reasoning.

3 Methodology

To address the challenge of knowledge-based reasoning in VQA, we propose a modular, multimodal architecture designed to effectively fuse information from visual, linguistic, and structured graph sources. Our methodology is built upon a carefully selected data foundation and features a so-

phisticated model architecture that leverages pre-trained foundation models, connected by specialized bridge components. The entire system is trained using a phased strategy to ensure stable and efficient learning.

3.1 Proposed Model Architecture

Our proposed architecture is not a monolithic network but a carefully orchestrated assembly of specialized, pre-trained encoders and a central language-based reasoner. The core design philosophy is to project heterogeneous data modalities into a common representational space that a Large Language Model (LLM) can interpret.

The process begins with the visual pathway, which employs a powerful, pre-trained vision encoder, specifically the Vision Transformer (ViT) from **CLIP (Contrastive Language-Image Pre-training)** (Radford et al., 2021). We choose CLIP’s ViT because its training on 400 million image-text pairs has endowed it with a rich semantic understanding of visual concepts that extends far beyond simple object recognition. To preserve this powerful, generalized knowledge and for computational efficiency, this vision encoder remains frozen during training. The encoder transforms the input image into a sequence of patch embeddings, which, while comprehensive, are too numerous and unstructured for direct LLM consumption. To bridge this gap, we introduce a **Querying Transformer (Q-Former)**, a critical component inspired by the **BLIP-2** architecture (Li et al., 2023). The Q-Former functions as an intelligent information bottleneck. It utilizes a small, fixed number of learnable queries to interact with the dense patch embeddings from the ViT, extracting only the most salient visual features relevant to the overall context. This process effectively distills a verbose visual representation into a compact set of feature vectors.

Concurrently, the structured knowledge pathway processes the input KG triples. Since knowledge graphs possess a distinct topological structure that sequential or convolutional models fail to capture, we employ a graph based decoder (Gardazi et al., 2025) which will extract relation based representation from a structured graph data. The GCN operates directly on the graph, iteratively aggregating feature information from neighboring nodes and edges. This message-passing mechanism produces node embeddings that are context-aware, resulting in a single, holistic **Graph Representation** vector.

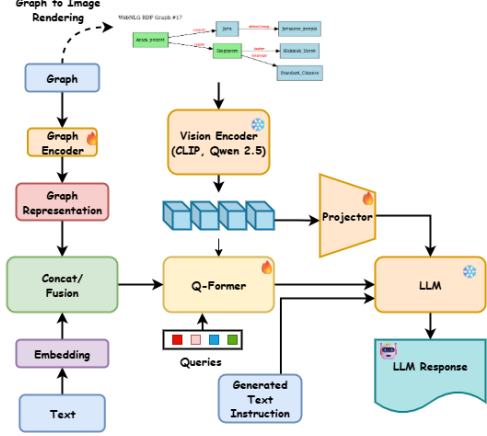


Figure 1: An overview of our proposed multimodal architecture for knowledge-augmented reasoning. The model integrates three distinct pathways (Graph, Text, and Vision) which are processed and fused before being interpreted by a Large Language Model (LLM). Components marked with a flame icon are trained, while those with a snowflake icon are kept frozen to preserve their pre-trained knowledge.

that encapsulates the relational knowledge of the entire KG snippet.

Finally, these disparate data streams are unified for the LLM. The input question is tokenized and embedded, then concatenated with the Graph Representation vector. The compact set of visual feature vectors from the Q-Former is passed through a lightweight, trainable **Projector**—a multi-layer perceptron (MLP)—which maps them into the same embedding space as the LLM’s vocabulary. The final input to the LLM is a sequence composed of these projected visual tokens (acting as a "soft visual prompt") followed by the embedded question and graph context. This unified sequence is then processed by a powerful, pre-trained LLM, such as **Llama 2** (Touvron et al., 2023) or **Mistral** (Jiang et al., 2023a), which serves as the central reasoning engine. By leveraging the LLM’s vast pre-trained knowledge and inferential capabilities, our model can reason over the provided multimodal context to generate a coherent, factually correct answer.

3.2 Training Strategy

Training such a complex, heterogeneous architecture requires a careful, phased approach to avoid instability and catastrophic forgetting. We therefore adopt a two-stage training strategy.

The first stage is dedicated to **Multimodal Alignment**. During this phase, both the large Vision Encoder and the LLM are kept completely frozen.

Training is focused exclusively on the bridge components: the Q-Former and the Projector. The objective is to teach these smaller modules the specific task of aligning the visual modality with the latent space of the LLM. By freezing the large backbones, we force the bridge components to learn how to extract and translate visual information into a format that the LLM can understand, without corrupting the powerful pre-trained weights of the larger models.

The second stage is **End-to-End Instruction Fine-Tuning**. Once the modalities are effectively aligned, we proceed to fine-tune the model for the specific VQA task. To make this process computationally tractable and to mitigate the risk of overfitting, we will employ **Parameter-Efficient Fine-Tuning (PEFT)**, specifically the **LoRA (Low-Rank Adaptation)** technique (Hu et al., 2021). LoRA avoids updating the millions of original LLM weights by injecting small, trainable rank-decomposition matrices into its transformer layers. This allows us to adapt the LLM’s behavior to our specific task and dataset by training only a fraction of the total parameters, significantly reducing the memory and computational footprint while achieving performance comparable to full fine-tuning.

3.3 Evaluation

The model’s performance will be quantitatively assessed using the standard VQA accuracy metric. To rigorously test the model’s reliance on the knowledge graph, we will curate a challenging split of the test set where the answers cannot be inferred from visual cues or general knowledge alone, thereby requiring explicit reasoning over the provided KG triples. Furthermore, we will conduct qualitative analyses by visualizing the attention or similarity maps generated within the Q-Former and ViT. This will allow us to interpret the model’s reasoning process by observing which parts of the image and graph it focuses on when formulating an answer, providing crucial insights into its "thought process."

4 Dataset Analysis

We analyze the dataset across three complementary views: natural text, node labels, and edge predicates. Table 1 summarizes the global statistics. Overall, the dataset comprises **38,872 texts** and approximately **760k tokens** in natural language, alongside **230k node labels** and **115k edge in-**

Metric	Value	Component Breakdown (Nodes / Edges)
Texts	38,872	–
Tokens	759,766	512,752 (Nodes) 117,557 (Edges)
Vocabulary	6,125	3,635 (Nodes) 412 (Edges)
TTR	0.0081	0.0071 / 0.0035
Rare words	1,046	64 / 7
Unique labels	–	3,624 (Nodes) 411 (Edges)

Sentence length statistics		
Average	19.5	–
Median	18	–
95th percentile	37	–

Table 1: Global statistics across modalities. TTR is the Type-Token Ratio. For brevity in the "Component Breakdown" column, some stats are presented in a *Nodes / Edges* format.

stances. The vocabulary sizes are **6,125** for text, **3,635** for nodes, and **412** for edges, yielding type-token ratios between 0.0035 and 0.008. Sentence lengths average 19.5 words, with a median of 18 and a 95th percentile of 37, showing moderate variability.

4.1 Part-of-Speech Distribution

POS distributions vary by modality. In natural text, nouns (**299k**), verbs (**132k**), adjectives (**81k**), and adpositions (**96k**) dominate, consistent with factual, entity-centric narratives. Node labels are heavily skewed toward common nouns (**383k**), adjectives (**67k**), and verbs (**21k**). Edges are almost exclusively noun-heavy (**114k**), with relatively few verbs or modifiers.

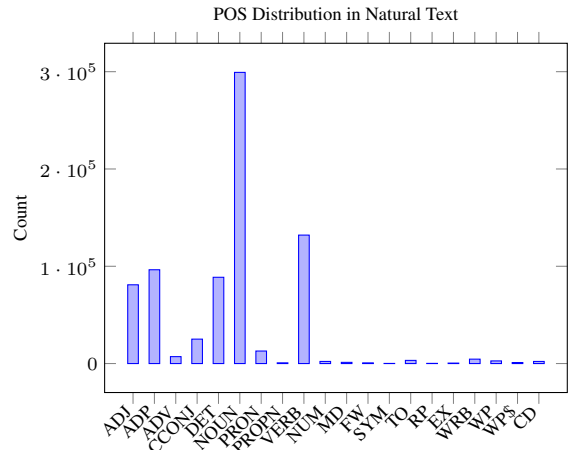


Figure 2: POS distribution in natural text.

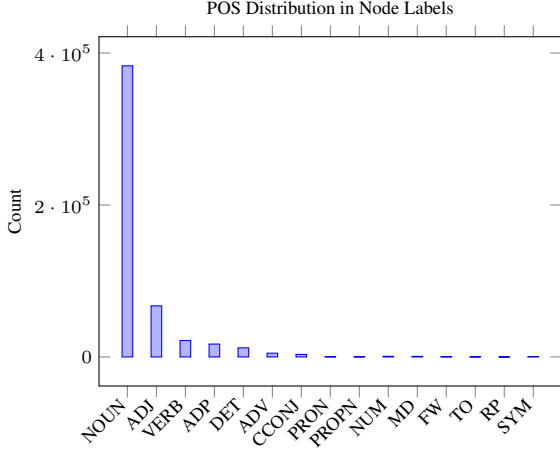


Figure 3: POS distribution in node labels.

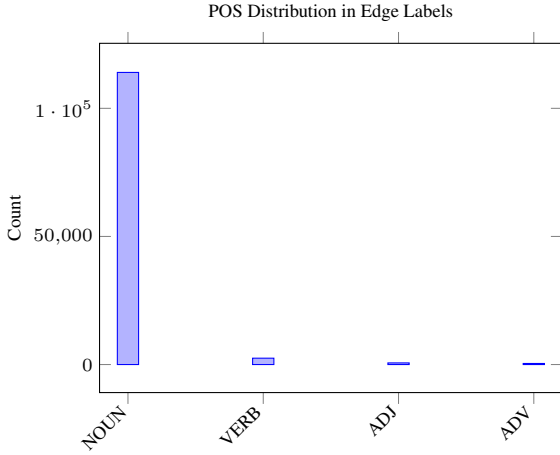


Figure 4: POS distribution in edge labels.

4.2 Cross-Modal Comparisons

To compare textual and structural components, we visualize vocabulary sizes, token counts, and rare word frequencies across modalities (Figure 5). Natural text is largest in raw token count, while node labels dominate in noun frequency. Edge predicates are narrower but dense in factual relations.

4.3 Implications

The dataset provides a balanced structure: natural text emphasizes verb/noun-rich descriptions, nodes supply compact entity labels, and edges encode relations with high noun dominance. This division supports multimodal learning: nodes/edges guide structured reasoning, while text offers naturalistic surface forms. The relatively low TTR values indicate manageable lexical diversity, while the presence of rare words ensures coverage of long-tail phenomena relevant for generalization.

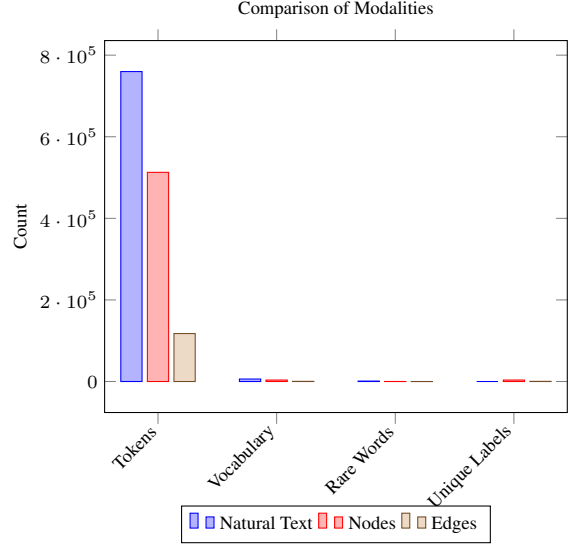


Figure 5: Cross-modal comparison of core statistics.

4.4 Implications for Our Task

The prevalence of proper nouns and factual predicates suits our goal of multimodal graph-based reasoning: aligned graph-text learning can leverage dense entity mentions, while the breadth of predicates (411 labels) encourages compositional generalization. The stable sentence lengths and modest vocabulary size simplify batching and optimization for sequence models, whereas the long-tail rare types motivate subword tokenization and label smoothing during training.

5 Empirical Analysis of Vision-Only Understanding

To motivate the necessity of a multimodal approach, we first conducted an empirical study to probe the capabilities and limitations of a state-of-the-art pre-trained vision encoder when tasked with interpreting a visually rendered knowledge graph. We processed an image of a graph from the WebNLG dataset (Figure 6a) using the Vision Transformer (ViT) from the CLIP model (Radford et al., 2021). By analyzing the model’s internal representations, we can form a baseline understanding of what a vision-only system perceives.

Our analysis, visualized in Figure 6, reveals that the ViT is highly effective at identifying regions of high visual saliency. The CLS vs. patch token similarity heatmap (Figure 6b) and its corresponding overlay (Figure 6a) clearly show that the model concentrates its attention on the graph’s nodes and the text they contain. Furthermore, the 2D PCA projection of the patch embeddings (Figure 6c)

Paradigm	Method	Core Approach	Key Innovation / Contribution	Primary Benchmark(s)
GNN / Embedding	EmbedKGQA (Saxena et al., 2020)	Embedding-Based Retrieval	Learns a joint embedding space for the question and the KG. Answers questions by finding the entity embedding closest to the question’s representation.	WebQSP, MetaQA
	RGCN (Schlichtkrull et al., 2017)	Graph Neural Network	A foundational GNN architecture that extends Graph Convolutional Networks to handle the multi-relational nature of knowledge graphs for link prediction and entity classification.	FB15k-237, WN18RR
Semantic Parsing	GrailQA (Gu et al.)	Seq2Seq Formal Query Generation	Translates natural language questions into a formal, executable query language (e.g., SPARQL). Achieves strong zero-shot generalization to unseen KG schemas and relations.	GrailQA, WebQSP
Large Language Model (LLM)	LLM + CoT (Wei et al., 2023)	In-Context Learning / Prompting	Leverages the emergent reasoning capabilities of massive LLMs (e.g., GPT-4) via Chain-of-Thought prompting to break down a complex question into logical steps and answer it without fine-tuning.	KQA Pro, ComplexWebQ
	Binder (Cheng et al., 2023) / StructGPT (Jiang et al., 2023b)	LLM with Tool Use (Fine-tuned)	Fine-tunes an LLM to serve as a reasoning agent that can generate and execute code (e.g., API calls, Python) to interact with various data sources, including knowledge graphs.	KQA Pro, Spider
	Our Proposed Model	Multimodal Fusion with LLM Reasoner	Uniquely fuses three modalities: structured graph representations (from a GNN), visual features (from a ViT + Q-Former), and text. Uses a frozen LLM as the central reasoning engine.	WebNLG-VQA (prop.)

Table 2: Comparison of State-of-the-Art (SOTA) paradigms for Graph-based Question Answering. The field has evolved from specialized embedding and GNN-based models to flexible, powerful systems that leverage Large Language Models (LLMs) as their core reasoning engine. Our proposed work aligns with the latest LLM-based trend but innovates by introducing a novel fusion mechanism for visual, textual, and structured graph data.

demonstrates a sophisticated ability to differentiate visual primitives; the model creates a tight cluster of tokens corresponding to the uniform white background, cleanly separating them from the more diverse tokens representing the graph’s content.

5.1 Limitations of a Purely Visual Approach

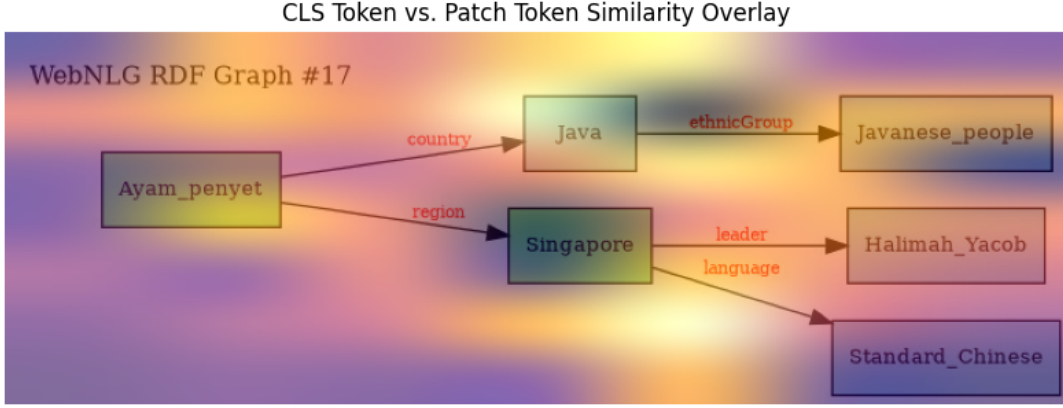
Despite this strong performance in visual feature extraction, this analysis exposes a fundamental deficiency: the model understands the graph’s **visual syntax** but fails to grasp its **semantic grammar**. The vision encoder perceives boxes, arrows, and text as collections of pixels, but it has no inherent mechanism to understand the abstract, relational concepts they represent. We identify three key lackings:

1. **Absence of Structural Awareness:** The ViT processes the input as a flat grid of patches. It has no native concept of "nodes" as distinct entities or "edges" as directed relationships that connect them. The arrow from ‘Singapore’ to ‘HalimahYacob’ is merely a set of dark pixels, indistinguishable in its structural role from the outline of a node box.

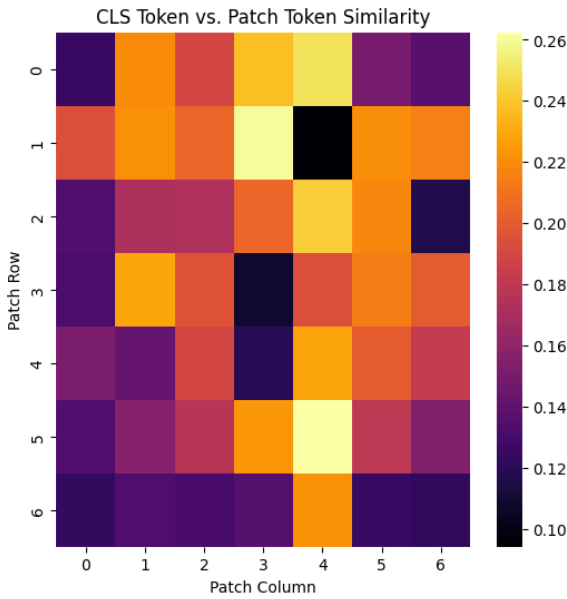
2. **Relational Ambiguity:** The model cannot semantically differentiate between edge types. The relationships ‘leader’ and ‘language’ are both rendered as red text. While a human uses context to understand their vastly different meanings, the vision model perceives them as visually similar patterns. It cannot infer that one signifies a person’s role while the other signifies a system of communication.

3. **Inability to Ground Entities:** The model sees the text "Singapore" but does not connect it to the symbolic entity `dbr:Singapore` in a knowledge base. This "semantic grounding" is a prerequisite for any form of true knowledge-based reasoning.

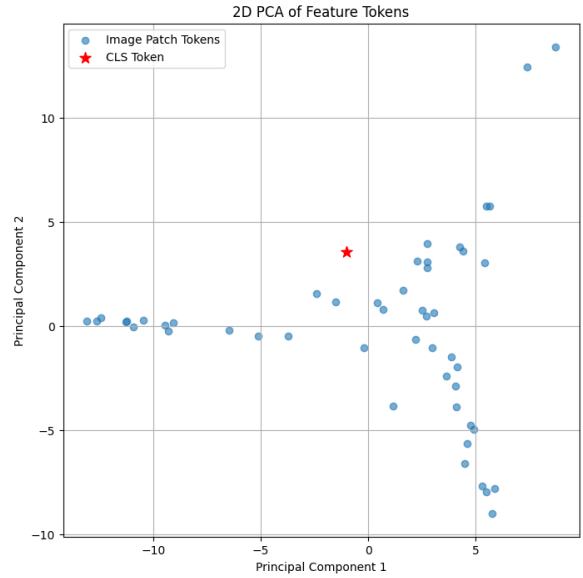
These limitations confirm that relying solely on a vision encoder, no matter how powerful, is insufficient for tasks that require deep relational understanding. The model provides a useful signal about *where* the important information is, but it cannot comprehend *what* that information actually means.



(a) Similarity heatmap overlaid on the original rendered graph.



(b) The raw 7x7 similarity heatmap.



(c) 2D PCA projection of the feature tokens.

Figure 6: Empirical analysis of a pre-trained Vision Transformer (ViT from CLIP) on a rendered knowledge graph. (a) The attention overlay shows the model focusing on visually salient regions like text and nodes. (b) The raw heatmap quantifies this focus. (c) The PCA plot demonstrates the model’s ability to cluster visual primitives (e.g., background vs. content), but reveals no inherent understanding of the graph’s semantic structure.

5.2 Mitigating Deficiencies with a Multimodal Architecture

Our proposed architecture is explicitly designed to overcome these lackings by treating each modality as a first-class citizen and fusing them intelligently.

First, to address the lack of structural awareness, we introduce a **dedicated Graph Encoder** (e.g., a GCN (Kipf and Welling, 2017)). This component operates not on the rendered image, but on the symbolic representation of the graph (its nodes, edges, and adjacency matrix). It produces a ‘Graph Representation’ that is unambiguous and topologically aware, directly encoding the relational facts that

the vision model fails to capture.

Second, to guide the vision system and bridge the gap between modalities, we employ a **Q-Former**, inspired by modern vision-language models (Li et al., 2023). The Q-Former acts as an intelligent information extractor. Instead of passively accepting all visual features, it uses a set of learnable queries—which are conditioned on the textual and graph context—to actively pull the most relevant visual information from the frozen ViT. This ensures that the visual signal is not just a generic feature map, but a targeted summary relevant to the specific reasoning task at hand.

Finally, by feeding the specialized outputs from

the Graph Encoder and the Q-Former into a powerful **Large Language Model (LLM)**, we enable true multimodal reasoning. The LLM receives a comprehensive, multi-faceted view of the problem: a symbolic understanding of the graph’s structure, a contextually relevant summary of its visual appearance, and the user’s textual question. This holistic input allows the LLM to perform the high-level inference that is impossible for any single-modality encoder alone.

References

- Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. *ICLR*, abs/2210.02875.
- Nadia Mushtaq Gardazi, Ali Daud, Muhammad Kamran Malik, Amal Bukhari, Tariq Alsahfi, and Bader Alshemaimri. 2025. [Bert applications in natural language processing: a review](#). *Artificial Intelligence Review*, 58(6):166.
- Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488. ACM.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023a. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023b. [StructGPT: A general framework for large language model to reason over structured data](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.
- Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. [Improving multi-hop question answering over knowledge graphs using knowledge base embeddings](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4498–4507, Online. Association for Computational Linguistics.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2017. [Modeling relational data with graph convolutional networks](#). *Preprint*, arXiv:1703.06103.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.