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 google/vit-base-patch16-224 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 extended Webnlg dataset to graph images and generated QA pairs for each graph

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.

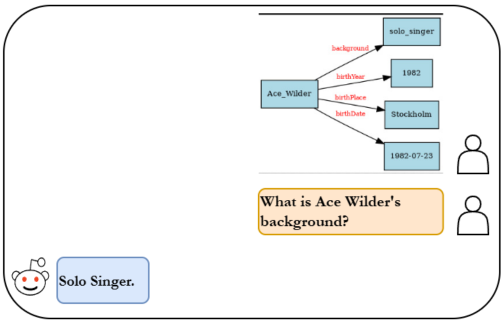


Figure 1: Example of our Graph QnA model

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 generalization 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., 2023) 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 Dataset Summary: Quad-Modal WebNLG Extension

This section summarizes the construction and characteristics of our extended **WebNLG dataset**, a benchmark for graph-to-text generation, transforming it into a **quad-modal resource** for multimodal reasoning.

3.1 Modalities and Extension

The base dataset aligns natural text with structured RDF triples, such as the example:

`(Aarhus_Airport,`
`elevationAboveTheSeaLevel, 25.0)`
Verbalizes to: "Aarhus Airport is 25.0 metres above the sea level."

The corpus was extended by introducing two additional modalities:

1. **Graph Images** (Tri-modal): Each graph is rendered into a high-resolution image, adding a visual/spatial modality.

2. **Question–Answer (QA) Pairs** (Quad-modal): $\mathbf{35,400}$ graphs were automatically augmented with $\mathbf{1}$ to $\mathbf{4}$ QA pairs per graph using the `gemini-2.5-flash-lite` API.

This results in a resource aligning $\textbf{Natural Text}$, $\textbf{Structured Graphs}$, $\textbf{Graph Images}$, and $\textbf{QA Pairs}$.

3.2 Illustrative Example (Table 1)

The tri-modal alignment is demonstrated in Table 1.

3.3 Dataset Statistics and Analysis (Table 2)

The corpus includes $\mathbf{38,872}$ text samples, $\sim \mathbf{760k}$ tokens, $\mathbf{230k}$ node labels, and $\mathbf{115k}$ edge instances. The vocabulary sizes are $\mathbf{6,125}$ (text), $\mathbf{3,635}$ (nodes), and $\mathbf{412}$ (edges). The average sentence length is $\mathbf{19.5}$ words (median $\mathbf{18}$, $\mathbf{95^{th}}$ percentile $\mathbf{37}$).

Table 2 summarizes the metrics:

Part-of-Speech analysis shows text is dominated by **nouns** ($\mathbf{299k}$), **verbs** ($\mathbf{132k}$), and **adjectives** ($\mathbf{81k}$), consistent with factual, entity-centric narratives and knowledge graph alignment.

3.4 Implications

The resulting quad-modal benchmark supports advanced tasks in graph-based multimodal reasoning, consistency checking, and cross-modal generation by aligning symbolic relations, linguistic grounding, spatial awareness, and evaluation-ready QA pairs.

4 Methodology: Multimodal Graph-to-Text Architecture

Our proposed approach, the **Multimodal Graph-to-Text Model**, is an architecture designed to fuse structured knowledge (graphs), visual information (images), and linguistic queries for coherent text generation. Conceptually based on the Querying Transformer (Q-Former), the model integrates features from an auxiliary Graph Encoder ($\mathcal{E}_G$) alongside a standard, frozen Vision Transformer (ViT).

4.1 Architectural Components

The architecture comprises three primary modules: the specialized Graph Encoder, a frozen Vision Encoder, and the trainable Q-Former Bridge ($\mathcal{T}_Q$), which connects these modalities to a frozen Language Model (LLM).

4.1.1 Graph Encoder ($\mathcal{E}_G$)

The $\mathcal{E}_G$ module converts the input graph structure into a dense vector representation. It consists of two stacked **Graph Attention Network (GAT)** layers (`GATConv`) that use self-attention to aggregate node features. Following node-level processing, an `AttentionalAggregation` layer performs global pooling to produce a context-aware graph embedding ($\mathbf{e}_G$). This feature is then mapped to the unified dimension D (typically 768) via a linear `self.graph_projection` layer: $\mathbf{e}_{G,\text{proj}} = \text{Linear}(\mathbf{e}_G)$.

4.1.2 Vision Encoder ($\mathcal{E}_V$)

A pre-trained **Vision Transformer (ViT)** serves as the $\mathcal{E}_V$. Its parameters are frozen during training to leverage general-purpose visual knowledge, yielding a sequence of visual feature vectors $\mathbf{V} \in R^{B \times S_V \times D}$.

4.1.3 Q-Former Bridge ($\mathcal{T}_Q$)

The Q-Former acts as an information bottleneck, transforming verbose multimodal features into a compact, fixed-length soft prompt for the LLM. It uses a fixed set of **learnable query embeddings** ($\mathbf{Q}$, e.g., 32 tokens) as input. Its stacked layers

Graph (Triple Form)	Text (Reference)	Graph QA (Example)	Graph Image
`Aarhus_Airport | elevationAboveTheSeaLevel | 25.0`	Aarhus Airport is 25.0 metres above the sea level.	**Q:** What is the elevation of Aarhus Airport above sea level? **A:** 25.0 metres	Aarhus_Airport elevationAboveTheSeaLevel 25.0

Table 1: Example instance showing multimodal alignment: symbolic graph triple, corresponding natural language description, generated QA pair, and the rendered graph image.

Metric	Value	Component Breakdown (Nodes / Edges)
Texts	38,872	–
Tokens	759,766	512,752 / 117,557
Vocabulary	6,125	3,635 / 412
TTR	0.0081	0.0071 / 0.0035
Rare words	1,046	64 / 7
Unique labels	–	3,624 / 411
Sentence length statistics		
Average	*19.5*	–
Median	*18*	–
95th percentile	*37*	–

Table 2: Global statistics across modalities in the extended WebNLG dataset.

Dataset	Number of QA Samples
Train	98114
Validation	4544
Test	5001

Table 3: Dataset sizes for QA samples.

refine these queries through three sequential steps: **Self-Attention** ($\mathbf{Q} \to \mathbf{Q}$), **Cross-Attention** ($\mathbf{Q} \to \mathbf{KV}$) over the combined multimodal input, and a **Feed-Forward Network (FFN)**. The output is a distilled sequence of multimodal tokens $\mathbf{Q}_{\text{out}} \in R^{B \times \text{num\_queries} \times D}$.

4.1.4 Language Model (LLM - Frozen)

A pre-trained sequence-to-sequence model (e.g., BART) is used for text generation, with its parameters entirely **frozen**. We access its embedding layer (`self.encoder_embedding`) only to convert the input question tokens into vector embeddings ($\mathbf{E}_{\text{ques}}$).

4.2 Forward Pass and Feature Fusion

The forward pass executes sequential feature processing and strategic fusion to construct the LLM's encoder input:

1. **Graph and Text Fusion:** $\mathcal{E}_G$ produces $\mathbf{e}_{G,\text{proj}}$. In the training phase, this is fused via mean-averaging with a projected embedding of the target text ($\mathbf{e}_{T,\text{proj}}$), if available, to create the fused graph-text sequence $\mathbf{S}_{GT}$.

2. **Multimodal Input Construction:** The frozen $\mathcal{E}_V$ produces $\mathbf{V}$. The combined features, $\mathbf{S}_{GT}$ and $\mathbf{V}$, are concatenated to form the Q-Former input $\mathbf{KV}$: $\mathbf{KV} = \text{Concatenate}(\mathbf{S}_{GT}, \mathbf{V})$.

3. **Q-Former Processing:** $\mathbf{KV}$ is processed by $\mathcal{T}_Q$ to yield the soft-prompt tokens $\mathbf{Q}_{\text{out}}$.

4. **LLM Input Construction:** The final encoder input $\mathbf{E}_{\text{final}}$ is constructed by prepending the soft prompt $\mathbf{Q}_{\text{out}}$ to the question embeddings $\mathbf{E}_{\text{ques}}$:

$$\mathbf{E}_{\text{final}} = \text{Concatenate}(\mathbf{Q}_{\text{out}}, \mathbf{E}_{\text{ques}})$$

During training, $\mathbf{E}_{\text{final}}$ is passed to the LLM encoder with the tokenized answers serving as decoder targets. For inference, $\mathbf{E}_{\text{final}}$ is used by the LLM's generation method.

5 Ablation Study: Direct Multimodal Fusion

To rigorously assess the contribution of the **Querying Transformer (Q-Former) Bridge** to cross-modal feature alignment and knowledge distillation, we implement a critical ablation study. We utilize a simplified baseline model that removes the Q-Former module entirely, replacing it with a straightforward linear fusion mechanism. This baseline is referred to as the **Direct Fusion Question Answering (DF-QA)** model.

5.1 Transfer Learning Context and Ablation Baseline

The DF-QA model is a direct structural predecessor to the main architecture. Before the QA fine-tuning, the components now frozen, specifically the Graph Encoder ($\mathcal{E}_G$), Vision Encoder ($\mathcal{E}_V$), and the initial linear fusion layer ($\mathcal{F}_{GV}$), were successfully utilized in a preliminary training phase. This phase involved fusing graph and vision embeddings to

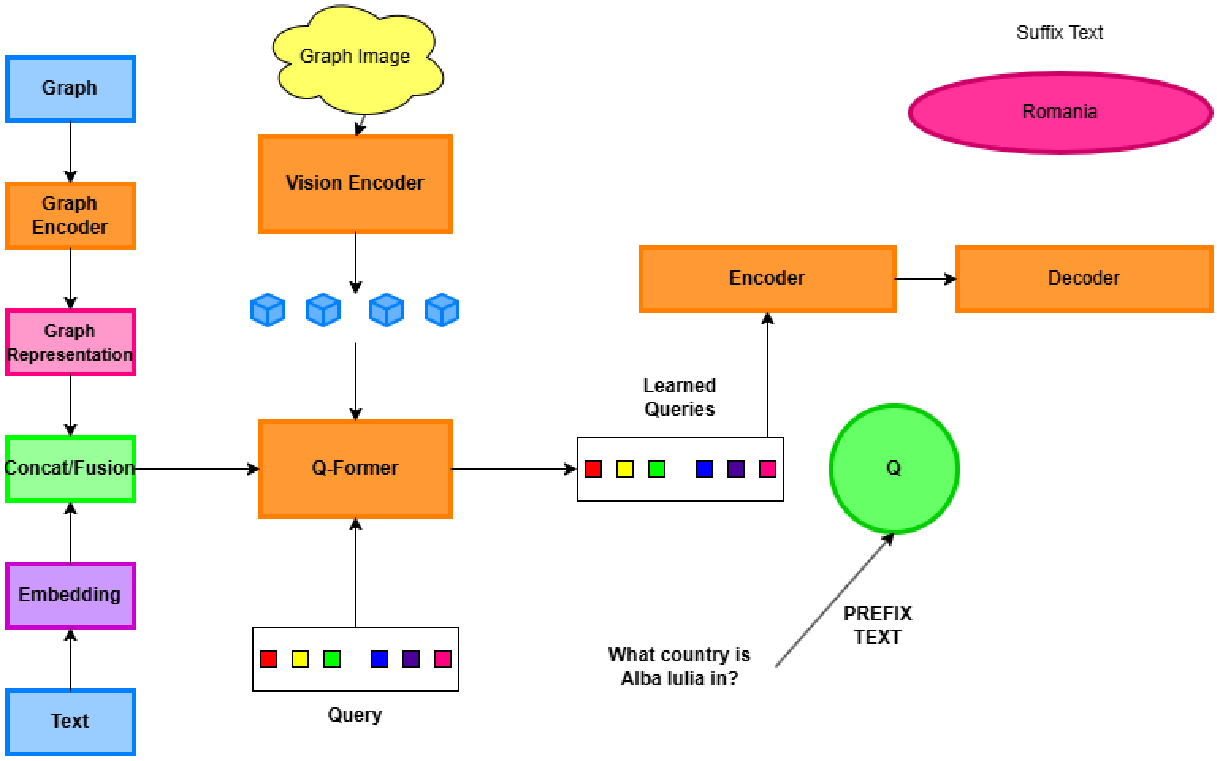


Figure 2: An overview of our KG-VQA pipeline. The model consumes three inputs: (i) a *graph* encoded into a symbolic *Graph Representation* via a Graph Encoder, (ii) the user *text question* embedded into a *Text Embedding*, and (iii) a *graph image* processed by a Vision Encoder to produce visual tokens. Text and graph representations are fused (*Concat/Fusion*). In parallel, a Q-Former attends over the visual tokens conditioned on the question tokens and emits a set of *learned query* vectors. The fused text–graph features together with the learned queries are provided to the LLM's Encoder, whose Decoder generates the final answer; the caption illustrates prefix conditioning on the question and a suffix example ("Romania").

generate natural language text (Graph-to-Text Generation), yielding promising initial results.

Consequently, the DF-QA model serves as a strong transfer learning baseline. Its weights for graph processing, vision processing, and the initial graph-vision fusion ($\mathcal{F}_{GV}$) are initialized from this successful pre-trained model. We then proceed to fine-tune this ablated structure on the Question Answering (QA) dataset to see the effect of missing the Q-Former layer in handling the specialized QA task.

5.2 Ablation Model Architecture: Direct Fusion QA (DF-QA)

The DF-QA model (`MultimodalGraphToTextQA`) replaces the iterative, attention-based knowledge extraction of the Q-Former with a two-stage linear fusion mechanism. This approach results in a highly constrained information path, collapsing all multimodal knowledge into a single vector.

5.2.1 Feature Extraction and Reduction (Frozen)

The initial feature extraction components remain frozen and operate identically to the main architecture:

1. **Graph Feature ($\mathbf{e}_G$):** The graph embedding from $\mathcal{E}_G$ is projected: $\mathbf{e}_G = \text{Linear}(\mathbf{e}_{\text{GNN}}) \in R^D$.

2. **Vision Feature ($\mathbf{e}_V$):** The ViT's `[CLS]` token embedding is extracted as the representative image feature $\mathbf{e}_V \in R^D$.

3. **Question Feature ($\mathbf{e}_Q$):** The input question tokens are embedded by the frozen LLM's embedding layer, and a mean-pooling operation is applied to create a single question vector $\mathbf{e}_Q \in R^D$.

5.2.2 Two-Stage Linear Fusion

The model employs two linear layers, both mapping concatenated inputs from $2D$ to D dimensions, followed by a $\tanh$ activation:

1. **Stage 1: Graph-Vision Fusion** ($\mathcal{F}_{GV}$) The two primary input modalities are concatenated and linearly fused. Crucially, the weights for this layer (`self.G_V_fusion`) are **transferred** from the initial Graph-to-Text pre-training task:

$$\mathbf{f}_{GV} = \tanh\left(\mathcal{F}_{GV}\left([\mathbf{e}_G; \mathbf{e}_V]\right)\right)$$

2. **Stage 2: Question Integration** ($\mathcal{F}_{Final}$) The resulting fused multimodal vector $\mathbf{f}_{GV}$ is then concatenated with the mean-pooled question feature $\mathbf{e}_Q$ and passed through a final linear layer (`self.final_fusion_layer`), which is trained during the QA fine-tuning:

$$\mathbf{f}_{\text{final}} = \tanh\left(\mathcal{F}_{\text{Final}}\left([\mathbf{f}_{GV}; \mathbf{e}_Q]\right)\right)$$

The final single vector $\mathbf{f}_{\text{final}} \in R^{1\times D}$ is unsqueezed and passed directly as the single input embedding (soft prompt) to the frozen LLM encoder.

5.2.3 Training Strategy

During the QA fine-tuning, all encoder and fusion component weights ($\mathcal{E}_G$, $\mathcal{E}_V$, $\mathbf{e}_{G,\text{proj}}$, $\mathcal{F}_{GV}$, and $\mathcal{F}_{\text{Final}}$) are **frozen**. Only the parameters of the LLM (BART) are fine-tuned. This procedure isolates the ablation to the structural difference, ensuring that the performance discrepancy is attributable solely to the missing Q-Former mechanism.

6 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.

7 Results and Discussion

Our experimental evaluation focused on training the primary Q-Former architecture and the Direct Fusion (DF-QA) ablation model on the Question Answering (QA) task. The results diverged significantly from our hypothesis, particularly regarding the performance of the Q-Former model.

7.1 Performance of the Q-Former Architecture (Main Model)

Contrary to expectations, the primary multimodal architecture incorporating the Q-Former Bridge exhibited **severe underperformance** on the generative QA task. Analysis of the training phase showed a critical failure in the learning process: both the training and validation loss curves remained largely **flat** throughout training. During inference, the model outputs were highly degraded, consisting mostly of **empty strings** or incoherent, garbled text (gibberish). This suggests that the Q-Former failed to generate a meaningful and useful soft prompt sequence, indicating that either the Q-Former's initialization or the specific token concatenation strategy with the question embeddings prevented the Language Model (LLM) from utilizing the input.

7.2 Performance of the Direct Fusion Ablation Model (DF-QA)

The simplified ablation model, Direct Fusion QA (DF-QA), which uses a single-vector linear fusion mechanism, unexpectedly yielded better results, generating human-readable and contextually relevant outputs, though its quantitative metrics remain low (Table 4).

Table 4: Generative Question Answering Evaluation Metrics for the DF-QA Ablation Model

Metric	Score
`N_Samples`	5001
QA Exact Match (EM)	0.049
QA F1 Score	0.0597
BLEU-4	0.0537
ROUGE-1 F1	0.0660
ROUGE-2 F1	0.0383
ROUGE-L F1	0.0657

The DF-QA model demonstrated a minimal capacity for knowledge retrieval, primarily due to a systematic generation bias. Predicted answers consistently contained content **related to the input graph structure**, confirming that the linear fusion successfully conveyed basic graph context. However, the model suffered from a **lack of question sensitivity**, generating the same answer for every distinct question associated with the identical graph. This indicates the single fused soft prompt ($\mathbf{f}_{\text{final}}$) primarily conveyed static graph/image context, overpowering the question embedding ($\mathbf{e}_Q$). Consequently, the low non-zero scores (EM, F1) are

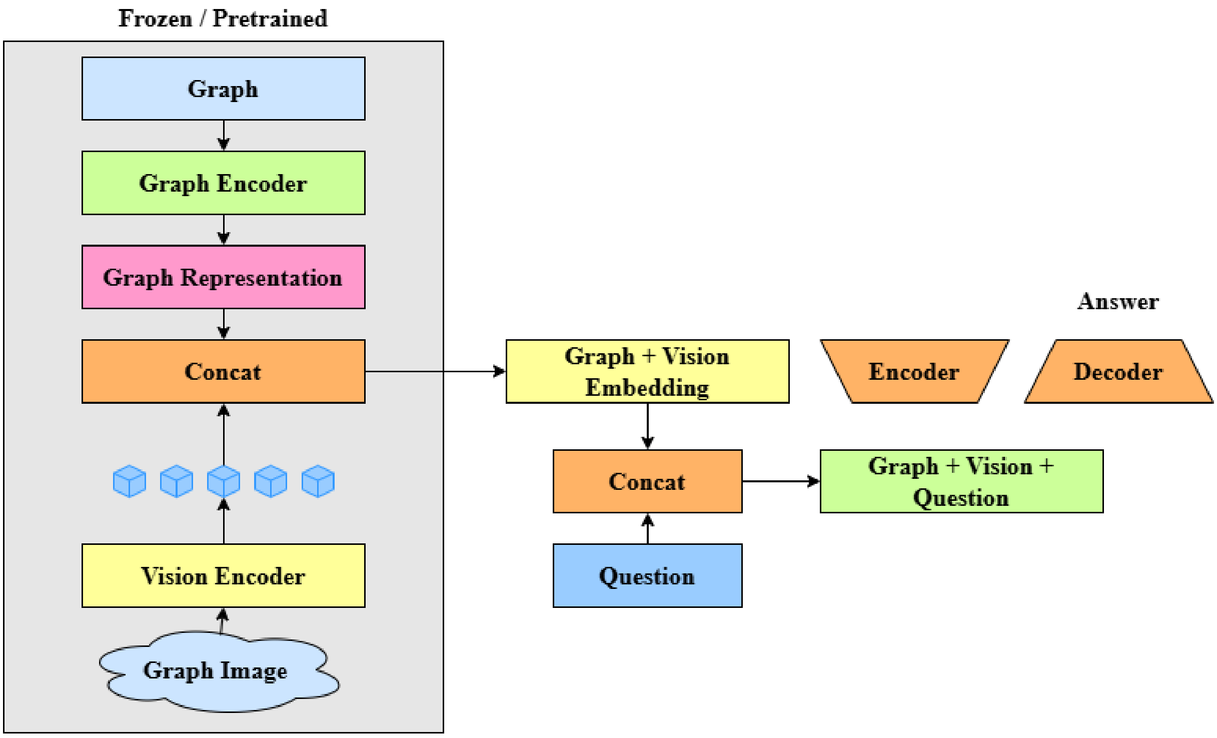


Figure 3: Ablation Architecture: Graph-Vision-Text Fusion Pipeline

attributed mainly to instances where this repetitive, graph-dependent output **coincidentally matched** the ground truth for some questions.

To make the qualitative results table shorter and more suitable for academic paper formatting, I will condense the triples and remove the "Graph/Image" column, placing the graph context above the table. This is a common practice when the same context is shared by multiple rows.

Here is the revised, condensed LaTeX table focusing on demonstrating the flaw of giving the same answer regardless of the question:

7.3 Conclusion of Ablation

The results suggest that the complexity introduced by the Q-Former structure hindered learning entirely in our setup. Conversely, the structural simplicity of the DF-QA model allowed the LLM to access and decode the basic graph context, albeit without the finesse required for question-conditioned generation. This unexpected outcome highlights that a simple, single-vector prompt was more effective than the failing multi-token prompt generated by the Q-Former. Further investigation into Q-Former initialization and optimization schedules is warranted.

8 Github Link

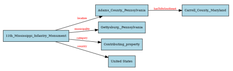
visit our github repo

9 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 pretrained vision encoder when tasked with interpreting a visually rendered knowledge graph. We processed an image of a graph from the WebNLG dataset (Figure 4a) 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 4, reveals that the ViT is highly effective at identifying regions of high visual saliency. The CLS vs. patch token similarity heatmap (Figure 4b) and its corresponding overlay (Figure 4a) 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 4c) demonstrates a sophisticated ability to differentiate visual primitives; the model creates a tight cluster of tokens corresponding to the uniform white

Table 5: Qualitative Results for the Direct Fusion QA (DF-QA) Ablation Model: Extreme Lack of Question Sensitivity

Visual/Graph Context (ID57)	**Input Question**	**Ground Truth Answer**	**DF-QA Predicted Answer**
Visual Reference:  Adams_County,_Pennsylvania hasToItsSoutheast Carroll_County,_Maryland location municipality Gettysburg,_Pennsylvania 11th_Mississippi_Infantry_Monument category Contributing_property country United States **Core Triples:** `Monument | location | Adams Co. Monument | municipality | Gettysburg Adams Co. | SE | Carroll Co. Monument | category | Property`	Where is the 11th Mississippi Infantry Monument located?	Adams County, Pennsylvania	**Adams County, Pennsylvania** *(Correct)*
	In which municipality is the 11th Mississippi Monument situated?	Gettysburg, Pennsylvania	**Adams County, Pennsylvania** *(Incorrect)*
	What is the category of the 11th Mississippi Infantry Monument?	Contributing property	**Adams County, Pennsylvania** *(Incorrect)*
	Which county is to the southeast of Adams County, Pennsylvania?	Carroll County, Maryland	**Adams County, Pennsylvania** *(Incorrect)*

Note: The DF-QA model incorrectly repeats the entity 'Adams County, Pennsylvania' for three of the four distinct questions. The single correct prediction is attributed to chance correlation (as it matches the model's static fixation), strongly indicating that the single-vector fusion $\mathbf{f}_{final}$ *acts as a static, question-independent graph summary, completely ignoring the input question* ($\mathbf{e}_Q$)*.*

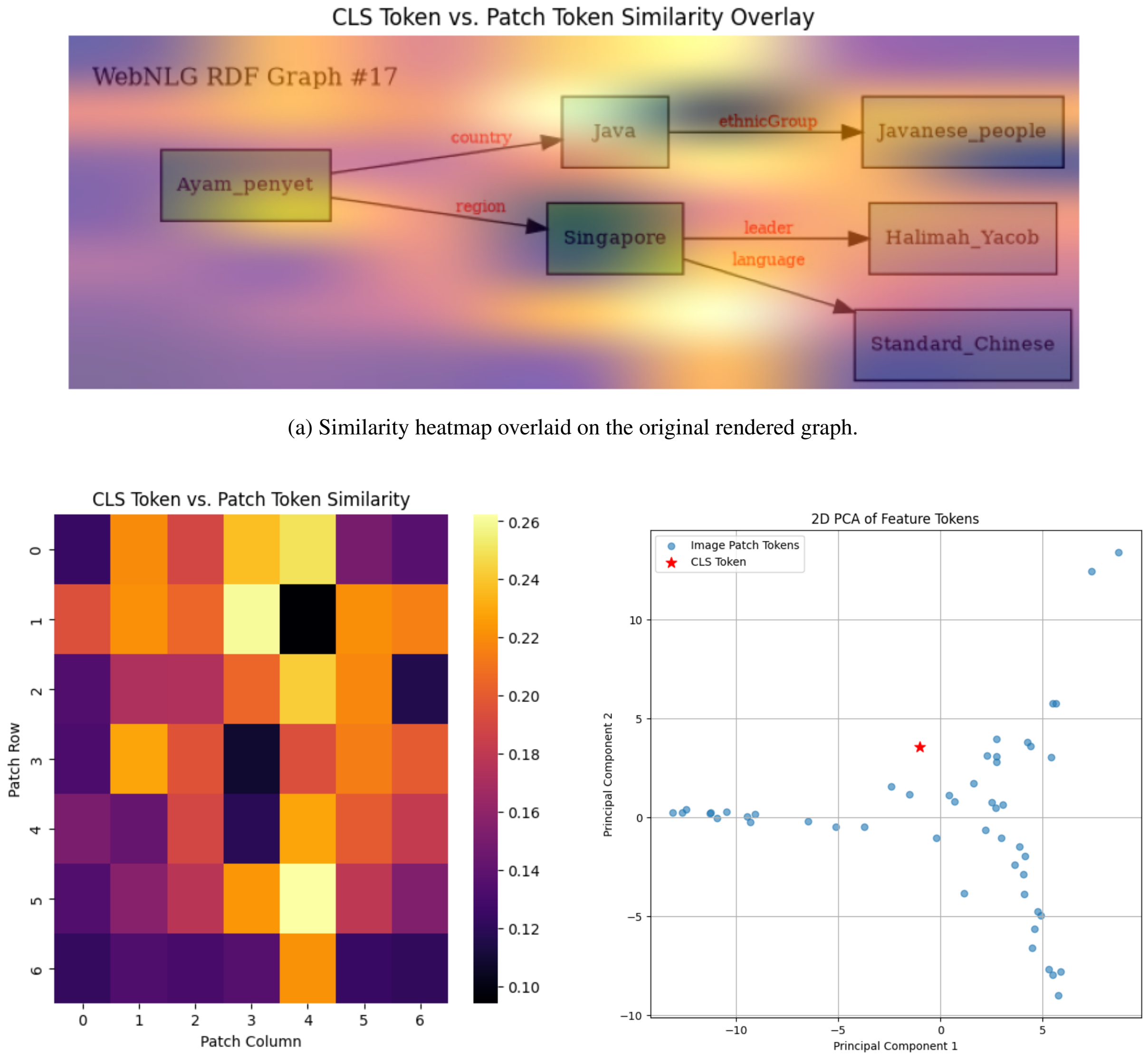
Paradigm	**Method**	**Core Approach**	**Key Innovation / Contribution**	**Primary Benchmark(s)**
GNN / Embedding	EmbedKGQA (Saxena et al., 2020)	Embedding-Based Retrieval	Learns a joint embedding space for the question and the KG. Answers questions by finding the entity embedding closest to the question's 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., 2023)	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 6: 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.

background, cleanly separating them from the more diverse tokens representing the graph's content.

9.1 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**



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

(b) The raw 7x7 similarity heatmap.

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

Figure 4: 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.

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.

9.2 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.

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.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. 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.

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.