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

Team: miniLTRC

Shubham Goel
shubham.goel@students.iiit.ac.in

Nidhi Vaidya
nidhi.vaidya@research.iiit.ac.in

Shree Mitra
shree.mitra@research.iiit.ac.in

August 10, 2025

Contents

1 Introduction

The goal of this project is to develop a **Multimodal Graph-Based Visual Question Answering (VGQA)** system. Using graph, image, and text modalities, it allows us to understand graphical data -such as airport relation triples (e.g., `"Aarhus Airport | cityServed | Aarhus, Denmark"`) - and generate descriptive summaries or answers.

By rendering the graph structure visually and combining embeddings from graph encoders, vision models, and text models, we aim to create a unified latent representation. This fused embedding will drive a large language model (LLM) to perform downstream tasks such as generating rich textual reports, answering questions, or summarizing graphs.

2 Methodology

2.1 Graph Preparation

- The rendered structured graph triples into a visual graph layout (e.g., via `NetworkX` and `Matplotlib`).
- Each sample comprises three modalities:
  1. **Graph structure**: Encoded by GraphFormer or other GNNs.
  2. **Rendered graph image**: Embedded via a vision encoder such as CLIP or Qwen 2.5 VL.
  3. **Original textual target**: Encoded separately using a text encoder.

2.2 Multimodal Embedding

- Extract a graph embedding from the graph encoder.
- Obtain patch-level image embeddings from the vision encoder.
- Get text embeddings from the text encoder (e.g., LLM's text embedding layer).

2.3 Fusion via Cross-Attention

- Fuse graph and image embeddings through a cross-attention module (e.g., Q-Former).
- Incorporate text embeddings by:
  1. Concatenating as another modality in the cross-attention stack, or
  2. Appending text tokens to the sequence of fused embeddings before sending to the attention block.
- Output: A single latent representation that spans graph, vision, and text.

2.4 Instruction & Projection

- Generate a synthetic instruction per sample, e.g., *”Write an extensive report on the given graph, including its content and context.”*

- Project each modality representation through MLPs to match the LLM input space.

2.5 LLM Interaction

- Feed the instruction, graph, image, and text representations into an LLM (e.g., Vicuna-7B, LLaMA).

- Train or fine-tune to produce comprehensive outputs for graph-to-text tasks (e.g., summarization, explanation, QA).

2.6 Evaluation

- Assess output quality using BLEU, ROUGE, and human evaluation.

- Benchmark against baselines that handle only single or fewer modalities.

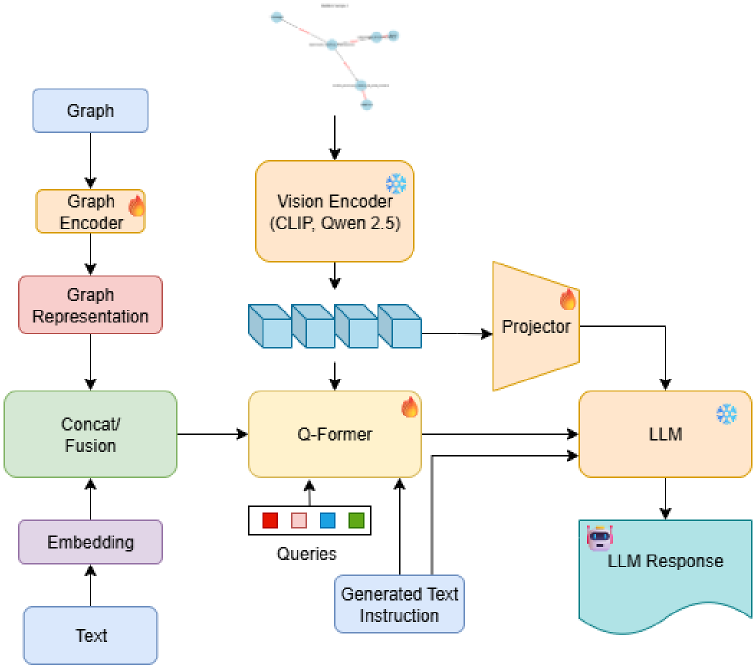

Figure 1: Overview of the proposed Multimodal Graph-Based Visual Question Answering (VGQA) framework. The system integrates structured graph representations, textual embeddings, and visual graph renderings through a fusion mechanism, followed by a Q-Former and projector to align multimodal features with a Large Language Model (LLM) for generating natural language responses.

3 State of the Art (Graph → Text & Multimodal Graph Learning)

Graph-MLLM A benchmark comparing GNN-based, LLM-based, and MLLM-as-Predictor approaches across multimodal graph tasks. In particular, MLLM-as-predictor methods, where the model itself generates graph output, achieve the best performance even without explicit graph structure encodings (arXiv).

UniGraph2 A cross-domain graph foundation model that fuses multimodal encodings (via modality-specific encoders plus a GNN), producing superior embeddings across multimodal graph benchmarks (arXiv).

GRAPHGPT-O A model that linearizes multimodal attributed graphs (with text and image nodes), aligns them via hierarchical encoders, and enables comprehension and generation in graph-based contexts (arXiv).

Graph4MM Integrates multi-hop graph structure into multimodal learning via hop-aware attention and a specialized cross-modal querying transformer (OpenReview).

These methods illustrate the successful fusion of graph structure with multimodal representations for generation or reasoning tasks.