# Project Report: Scalable Multimodal Image Captioning via Q-Former and DistilGPT-2

Thomas Reedy,
Ahlad Sajjanam,
Jagath Kumar Reddy Katama Reddy

December 2025

## 1 Introduction

This project focused on building a multi-modal image captioning system that generates descriptive language conditioned on image content. We implemented a transformer-based architecture inspired by recent state-of-the-art models such as BLIP-2. The model combines a Vision Transformer (ViT) for visual feature extraction with a DistilGPT-2 decoder for caption generation. A Q-Former module bridges the modalities by selecting and projecting relevant visual tokens into the language model embedding space.

## 2 Methodology

To achieve robust multi-modal generation, we implemented a modular transformer-based architecture inspired by the BLIP-2 framework. Our approach moves beyond simple concatenation of features, utilizing a **Querying Transformer (Q-Former)** to bridge the modality gap.

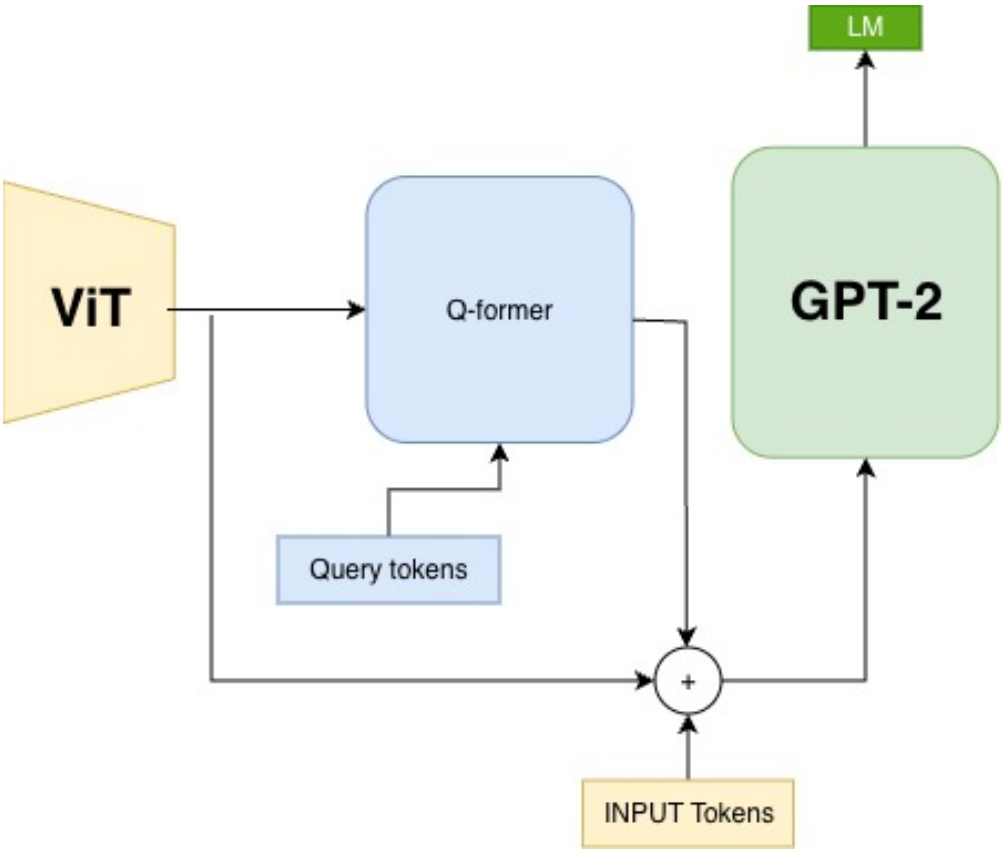


Figure 1: This is our high level representations of the model architecture. Frozen ViT. Q-former with 16 learnable query tokens. Distill-GPT-2 (82M parameters) with final 4 layers unfrozen. Image feature + q-former prompts +¡start¿ are feed into GPT module as prefix tokens. The objective function is same as Language modeling-cross entropy loss

## 2.1 Visual Encoding (ViT)

We utilized a **Vision Transformer (ViT)** as our backbone encoder. Unlike traditional CNNs, the ViT processes the image as a sequence of patches, extracting global context embeddings. This ensures that the model captures long-range dependencies within the image structure immediately.

## 2.2 The Bridge: Q-Former

The critical innovation inspired from BLIP-2 is the **Q-Former**. Rather than passing the raw ViT embeddings straight into the language model, the Q-Former works as a smart, trainable middle layer. It uses a fixed set of learned visual queries (we used 16—half of what BLIP-2 uses) to pull out only the most meaningful parts of the image representation. Those selected features are then transformed into a form that fits naturally into the language model's embedding space. In practice, this lets the visual signals line up much better with the decoder's semantics, making the whole vision-to-language pipeline more coherent and efficient.

### 2.3 Semantic Adaptation (DistilGPT-2)

For the decoder, we employed **DistilGPT-2**. We chose this distilled version to reduce net training time and computational overhead without significantly sacrificing language fluency. **Partial Unfreezing Strategy:** To ensure the model actually "looks" at the image rather than just relying on its pretrained language knowledge, we **unfroze the last four layers** of the decoder. This allowed for semantic adaptation, enabling the model to adjust its generation based on the visual context provided by the Q-Former.

## 3 Experiments

We designed our experiments to validate our architecture (Proof-of-Concept) and its scalability to complex data.

### 3.1 Experimental Design

We conducted two distinct phases of experimentation:

1. **Phase I: Flickr8k (Proof-of-Concept)**: We first trained on the smaller Flickr8k dataset. This allowed for rapid iteration and hyperparameter tuning to validate that the Q-Former was correctly bridging the modalities. Specifically, this was trained on 6k training images and 1k validation images together over 10 epochs.

2. **Phase II: MS-COCO (Scalability Benchmark)**: When validation of the architecture, we scaled the training to the MS-COCO dataset (train2014). This experiment was designed to test the model's ability to generalize and produce richer, more diverse descriptions. Specifically, this was trained on 414k Image-captions pairs (over 82k images) over 12 epochs. It took more than 20 hours to run on a single node of the A100 GPU.

## 4 Results

Our experiments demonstrated potential convergence, with training loss consistently decreasing across epochs, indicating effective learning of vision-language mappings.

### 4.1 Quantitative Analysis

The model achieved moderate performance on standard metrics (CIDEr-4 and Rouge-L), validating that the unfreezing of the DistilGPT-2 layers works on alignment but far from state of the art. Following are the metrics (Fig-2):

| Model | CIDEr | Rouge-L |
|---|---|---|
| X-LAN | 120.7 | 57.6 |
| BLIP* | 113.0 | 41.9 |
| Our Model (COCO)** | 56.1 | 24.4 |
| Our Model (flickr)** | 25.7 | 16.0 |

* referenced from official BLIP paper
**evaluated on COCO validation split

Figure 2: Comparison of image-captioning performance across models using CIDEr and Rouge-L scores, showing the performance gap between baseline models (X-LAN, BLIP) and our COCO/Flickr-trained models

## 4.2 Qualitative Findings

The system demonstrated significant capabilities in **Visual Reasoning**: **Object Awareness:** The model successfully identifies primary objects (e.g., "dog," "frisbee," "car").

**Vocabulary Improvement:** The larger MS-COCO dataset allowed the model to use more accurate and relevant adjectives compared to the Flickr8k model.

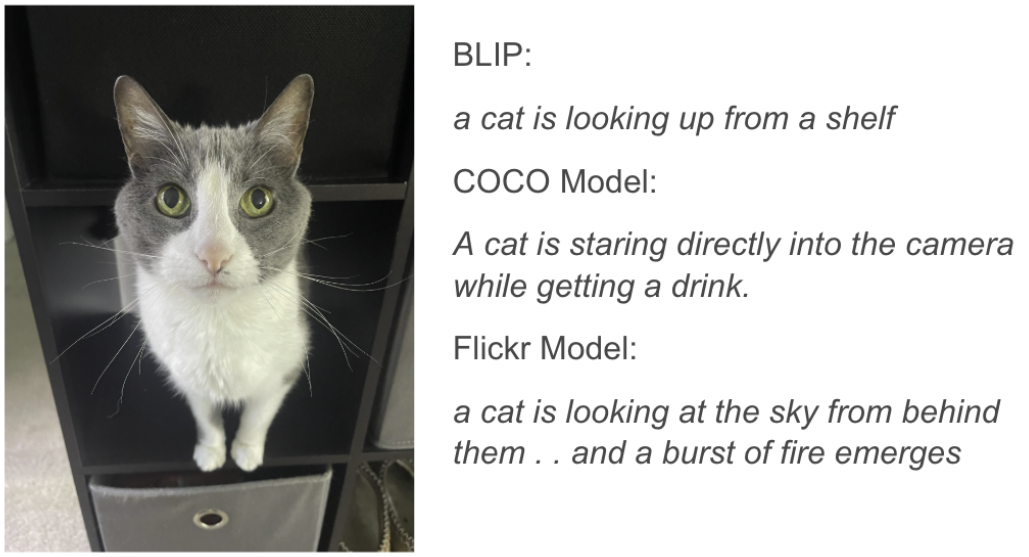


Figure 3: Captions generated by BLIP, our COCO-trained model, and our Flickr-trained model for an input cat image, highlighting qualitative matching in description accuracy

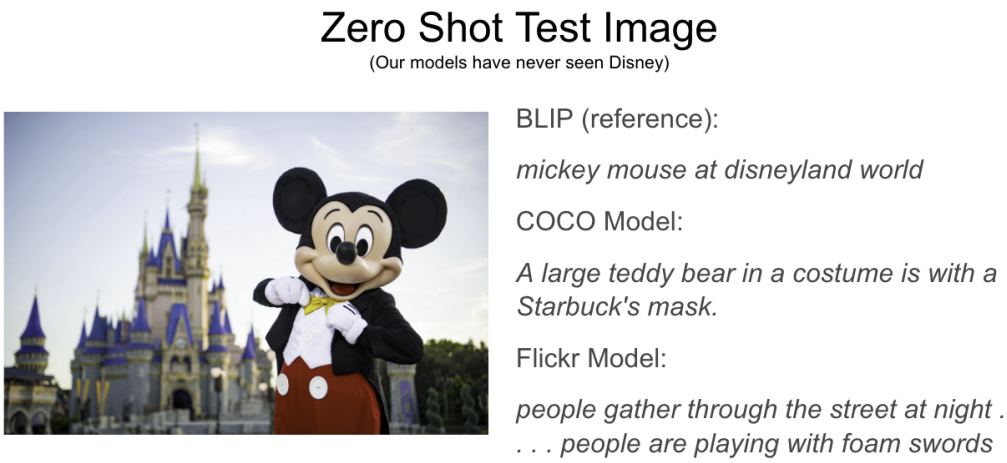


Figure 4: Captions generated by BLIP, our COCO-trained model, and our Flickr-trained model for an input of Mickey Mouse which our model has not seen before,highlighting the generation efficiency

### 4.3 Known Failure Cases & Limitations

Despite the architectural successes, our analysis revealed specific failure modes common in models lacking massive-scale pre-training (such as CLIP-level scale):

**Hallucination:** On the smaller Flickr8k dataset, the model occasionally generated objects not present in the image.

**Inconsistency:** When model is run on the same image multiple times to generate multiple captions samples, the model's output are fairly inconsistent which is clearly visible from Figure-6.


Figure 5: Ground-truth versus generated caption (Model trained on flickr) for a girl in a white dress, demonstrates hallucinations made by our model when interpreting objects and actions in real-world photos


Figure 6: This is one of the examples, where Model trained on Flickr8k dataset fails at being consistent about its understand of the images: CAPTION-1:"as seen from behind a black car . Sound waves of a man in a brown hoodie . Samsung Electronics"; CAPTION-2:"men on hoodies run down a dirt road . colorful graffiti . larger picture shows 3 women riding . blurry pictures . larger picture shows a bikini on"; CAPTION-3:"2nd St wetsuit owner helps someone ride a car . . . out of a wooded area . . with trees and some surrounding dirt"