Adapting BLIP2 for Visual Abductive Reasoning

Vedant Bhasin, Sayali Kandarkar, Qin Wang

Carnegie Mellon University, Language Technologies Institute

ABSTRACT

Vision-Language Models (VLMs) have significantly advanced visual content interpretation, but their abductive reasoning is limited in scenarios that require more fine grained reasoning. In this work, we explore on the Sherlock dataset, trying to improve VLMs—including CLIP RN50x64-multitask, and BLIP2 for the task of abductive reasoning. The main method we tried is to use a generated scene graph to give additional clues for the model, so that the model has a comprehensive and fine-grained understanding of the context of the whole image.

Our study aims to bridge the gap between machine learning models and human-level abductive reasoning, providing a pathway to more resilient and generalizable VLMs.

MOTIVATION

Abductive reasoning, a crucial element in AI applications like natural language understanding and computer vision, is often challenging for VQA models to navigate, particularly in nuanced context. Our focus is on the Sherlock dataset, a unique blend of textual prompts and images designed to enhance abductive reasoning. Comprising 363K inferences grounded in 103K images, each annotated with meticulous object bounding boxes, we set out to unravel the complexities of abductive reasoning. Our research is centered on employing scene graph generation to illuminate fine-grained relationships between objects in an image. Join us in this exploration, where we aim to advance the understanding of abductive reasoning, ultimately refining AI capabilities for more nuanced and context-aware applications.

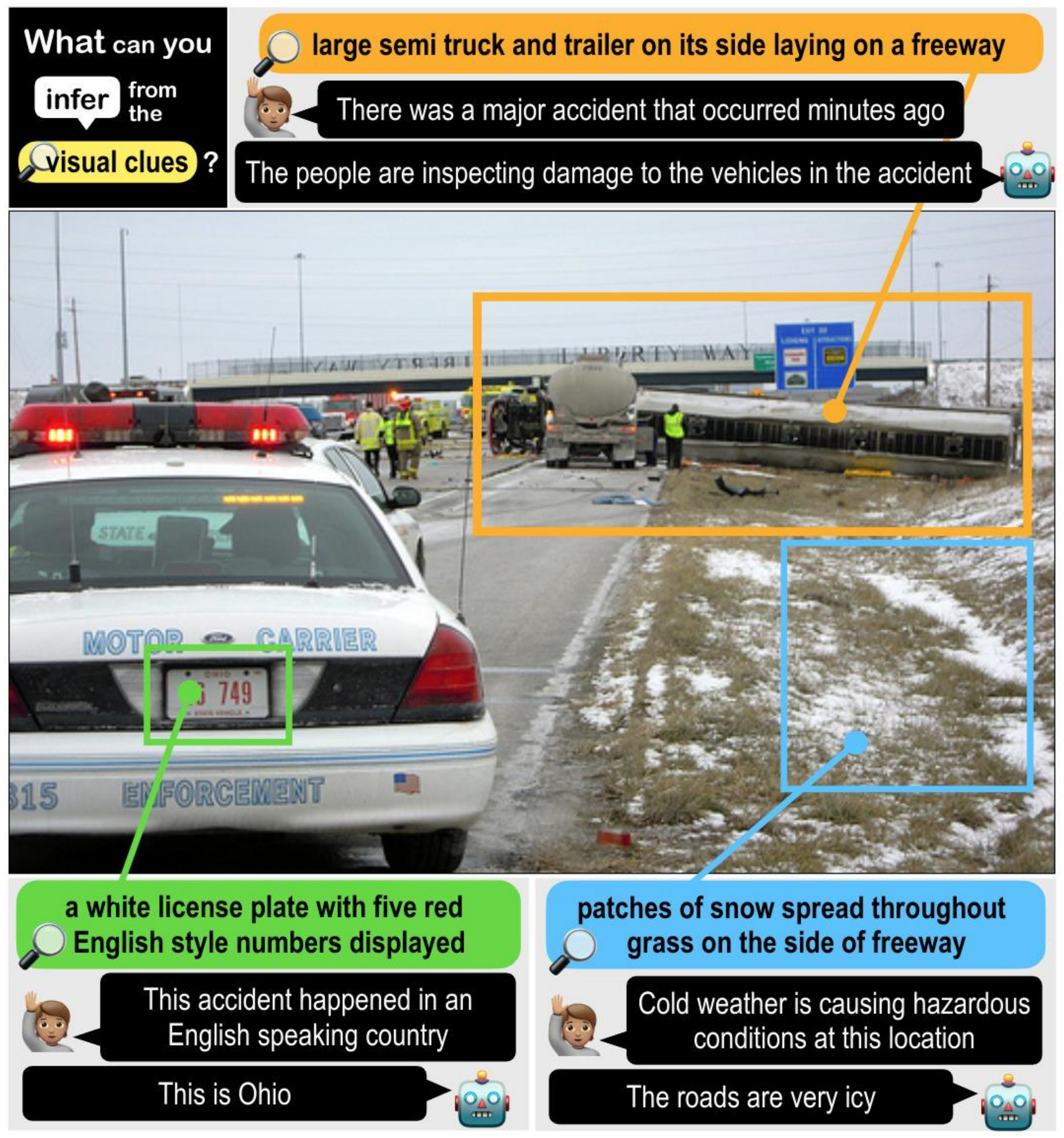
What can you infer from the visual clues ?
large semi truck and trailer on its side laying on a freeway
There was a major accident that occurred minutes ago
The people are inspecting damage to the vehicles in the accident
a white license plate with five red English style numbers displayed
This accident happened in an English speaking country
This is Ohio
patches of snow spread throughout grass on the side of freeway
Cold weather is causing hazardous conditions at this location
The roads are very icy

Figure 1. Sherlock Dataset Representation

PROPOSED METHODS

Scene Graph Generation

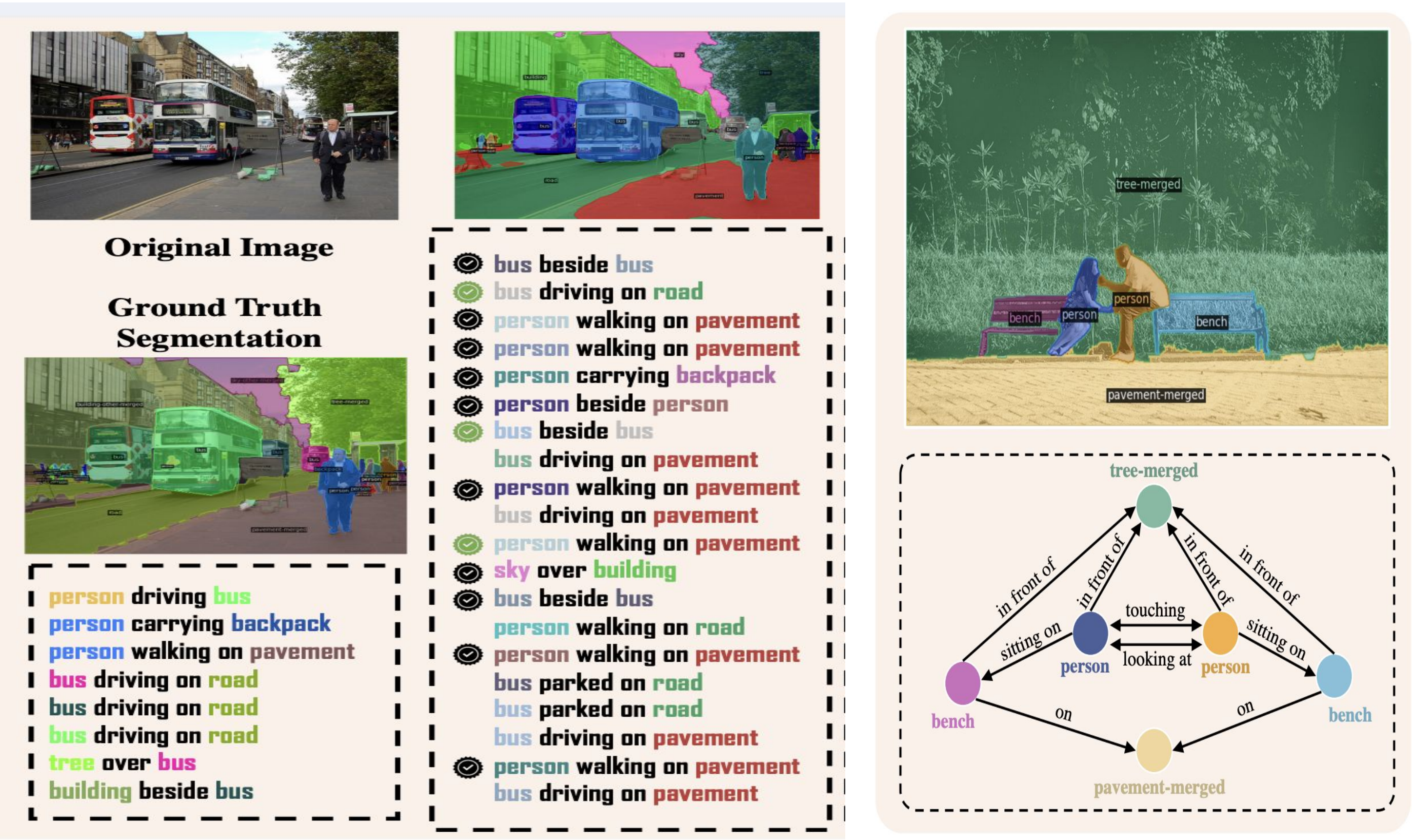

Figure 2. Left: Segmentation after Scene Graph Generation; Right: Graphical Representation

To capture fine-grained relationships between objects in an image, we used a Panoptic Scene Graph model, PSGFormer (J. Yang et al., 2022) to generate scene graph for each image. Given an input: (image, bounding box(es)), for each segmentation region in the scene graph, we calculated the percentage of area that is contained in the bounding box, if greater than 50%, we add the relevant triplet [subject, predicate, object] to the relationships. These relationships are cataloged in textual format and subsequently integrated into multimodal models alongside the image data. This combined input empowers the generation of fused features, enriching the representation of the input data across modalities.

TRAINING METHODOLOGY

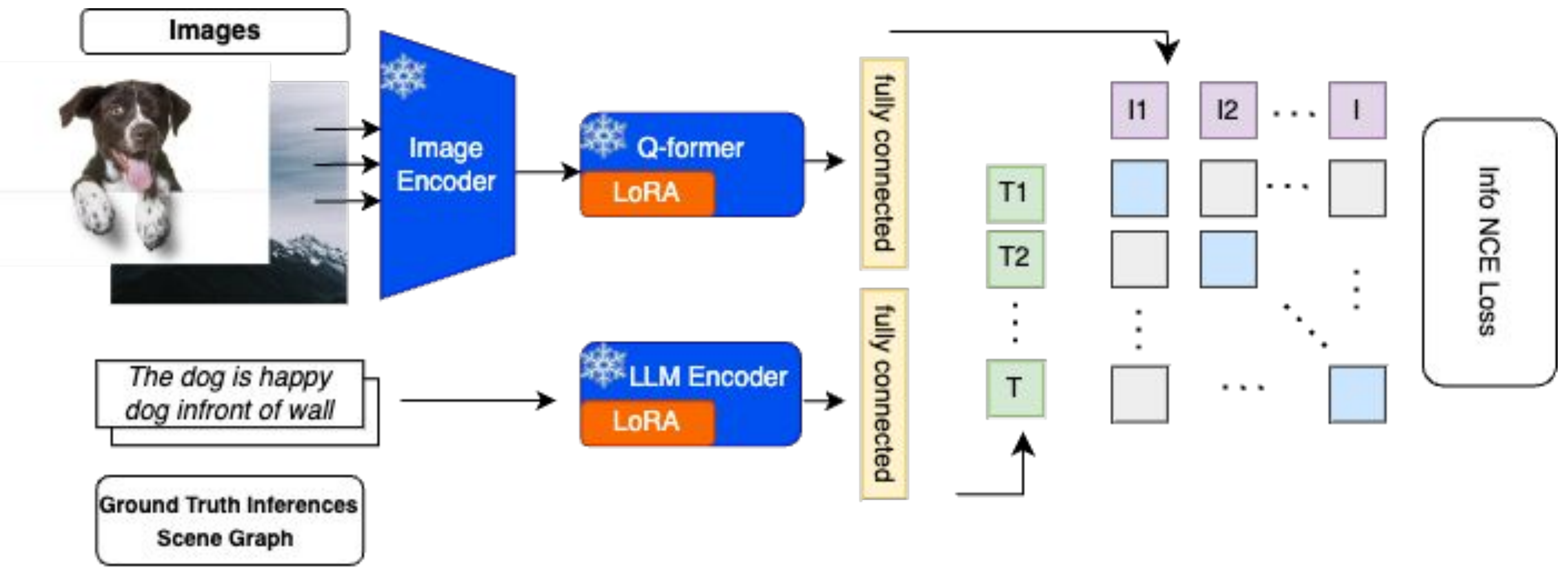

Figure 3. BLIP2 Fine-tuning pipeline using LoRA adapter

We inject LoRA adapters (Hu, Edward J., et al.) into the query and key projection layers of the Q former and the Encoder of the LLM. We obtain the image embedding by passing the image through the image encoder and the Q former, and we obtain the text embedding by passing in the ground truth inference and a textual representation of the scene graph through the encoder of the BLIP 2 (Li, Junnan, et al.) language model. Similar to CLIP, once the image and text embeddings are obtained, the model is trained to predict which image - inference + scene graph pair match using InfoNCE Loss (Oord, et al)

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log\frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

RESULTS

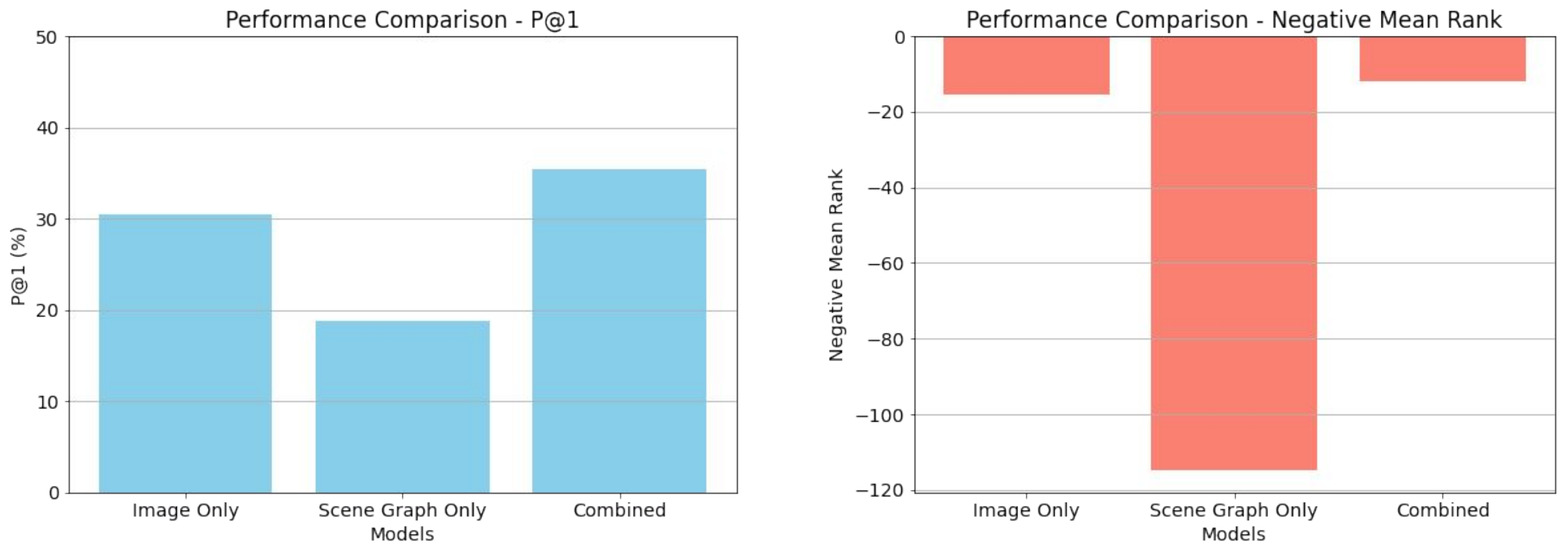

Figure 4. Left: the Precision at 1 (P@1↑) scores for three distinct models—Image Only, Scene Graph Only, and Combined. Right: the negative mean rank (↑) for three models. For both images, the task is: given image with bounding box, retrieve the best inference from 1000 candidate inferences.

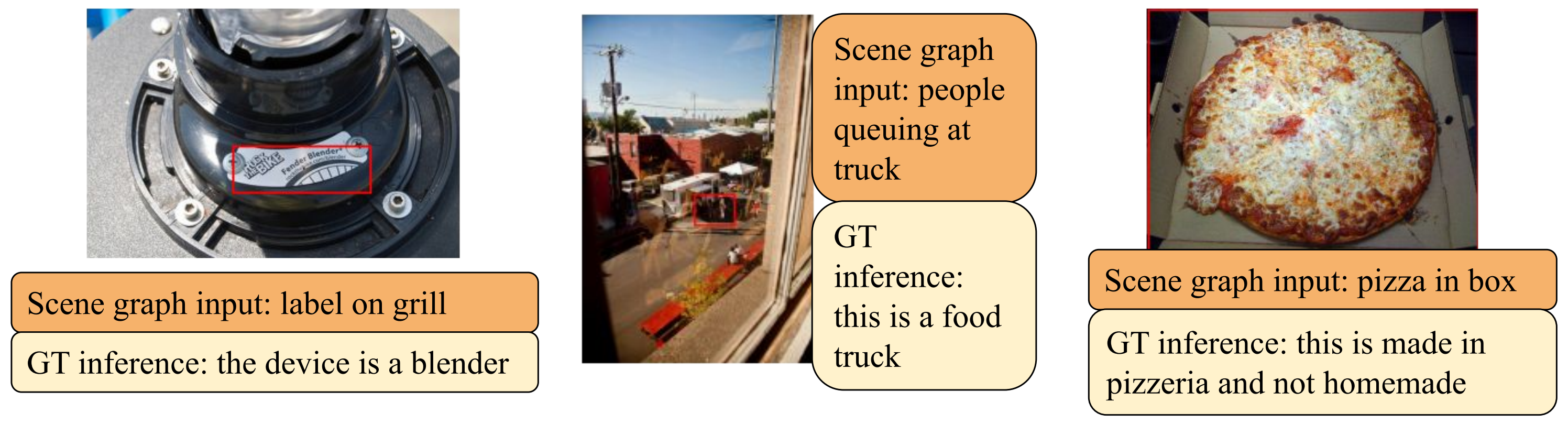
Scene graph input: label on grill
GT inference: the device is a blender
Scene graph input: people queuing at truck
GT inference: this is a food truck
Scene graph input: pizza in box
GT inference: this is made in pizzeria and not homemade

Figure 5. Left: example which image-only model retrieved correctly but not scene-graph-only model; Middle: example which scene-graph-only-model retrieved correctly but not image-only model; right: example which combined-model retrieved correctly but not uni-modality models;

P@1 (Precision at 1):

- The P@1 for the combined model (34.4%) outperforms both the image-only model (30.5%) and the scene graph-only model (18.8%). This suggests that incorporating both image and scene graph information enhances the model's ability to accurately predict the top-ranked output.
- The substantial increase in P@1 from the scene graph-only model to the combined model signifies the complementary nature of these modalities in improving prediction precision.

Negative Mean Rank:

- The negative mean rank indicates the negative average rank of the correct prediction among the outputs, where higher values denote better performance.
- Notably, the combined model exhibits the lowest negative mean rank (-13.12), implying that, on average, the correct predictions rank higher when both image and scene graph information are utilized. This indicates the combined model's ability to more consistently rank the correct predictions higher compared to the other models.

CONCLUSIONS

In summary, the adaptation of the BLIP2 model for Visual Abductive Reasoning demonstrates significant advancements in AI's ability to interpret complex visual content in conjunction with textual data. Our experiments on the Sherlock dataset showcase that incorporating scene graph generation into the model enhances its fine-grained relationship understanding between objects in an image. The results indicate a substantial improvement in the Precision at 1 (P@1) scores, especially when combining both image and scene graph information compared to using each modality individually. This hybrid approach not only improves top-ranked inference prediction but also exhibits a lower negative mean rank, suggesting more consistent accuracy. These findings underline the potential of context-aware representations in refining AI capabilities for more nuanced and robust visual reasoning tasks, paving the way for further research into multi-modal AI systems that can perform at near-human levels of abductive reasoning