
Fine Grained Image Grounding for Visual Abductive Reasoning

Vedant Bhasin^{* 1} Sayali Kandarkar^{* 1} Qin Wang^{* 1}

Abstract

In this work, we address the challenge of fine-grained image grounding for Visual Abductive Reasoning (VAR). Leveraging the Sherlock dataset, we focus on enhancing the interpretative capabilities of Vision-Language Models (VLMs) by integrating panoptic scene graphs (PSG) generated based on image bounding boxes. Our approach involves fine-tuning BLIP-2 with PSG, enriching the model’s understanding of intricate visual contexts. Our findings demonstrate improvements in precision at 1 (by 2.6%) and negative mean rank (by 3.3) with the inclusion of scene graphs and images over just image, indicating the efficacy of our method in enabling VLM’s to interpret complex visual content with greater accuracy and depth.

1. Introduction

Abductive reasoning, the ability to seek the simplest and most likely conclusion from a set of observations, plays a pivotal role in various real-world applications such as natural language understanding, computer vision, and artificial intelligence. Understanding the nuances of human abductive reasoning in the context of multimodal data is crucial for advancing these technologies. Recent study (Chen et al., 2023) shows that although many state of art VQA models are very good at extracting the surface level information when given a question, many of them fail to find the nuances in the context in order to give users more interesting answers. In this study, we delve into the Sherlock dataset (Hessel et al., 2022), a rich resource that challenges the boundaries of abductive reasoning by combining textual clues with visual information. This problem is important for the research community because it represents a significant step towards replicating human-like capabilities in going beyond visual

recognition, thereby enhancing the decision-making capabilities of automated systems in complex, real-world scenarios.

We address the challenge of explicitly encoding fine-grained information about the inference target region. Previous work in the area of fine-grained image grounding for VAR focuses on prompt tuning techniques such as Colorful Prompt Tuning (CPT) (Yao et al., 2022) and Regional Prompt Tuning (RGP) (Zhang & Fernando, 2023) both of which work on mechanisms through which the image embedding is generated from the CLIP encoder. To the best of authors’ knowledge, grounding the image with inputs exclusively to the text modality in VAR tasks has not been explored.

In this paper, we use a panoptic scene graph generator (Yang et al., 2022) to get a scene graph for the target inference region of the image. We represent this scene graph in a language format as a series of triplets of object predicate object. This is illustrated in Figure 1. This approach has several advantages. Firstly, encoding the scene graph information in language allows for easy integration of this information with any language model without any architectural changes as would be required for other representations (GNN’s for example). Secondly, this approach is highly modular allowing for different components such as the image encoder, language model, or scene graph generator to be swapped out for different models with minimal changes. Finally, it requires no augmentations on the image data working exclusively with the language modality. Our approach centers on using a contrastive loss to generate image embeddings which are strongly correlated with the text embeddings of the ground truth inference and the bounding box scene graph.

Our main contribution includes pioneering the explicit encoding of fine-grained details about the inference target region in Visual Abductive Reasoning (VAR) tasks. Unlike existing methods relying solely on the CLIP encoder for image embedding generation, we introduce a novel approach that grounds images using inputs exclusively from the text modality, thereby expanding the scope of multimodal reasoning in VAR.

^{*}Equal contribution ¹Carnegie Mellon University, Pittsburgh, PA. Correspondence to: Vedant Bhasin <vedantb@andrew.cmu.edu>, Sayali Kandarkar <skandark@andrew.cmu.edu>, Qin Wang <qinw@andrew.cmu.edu>.

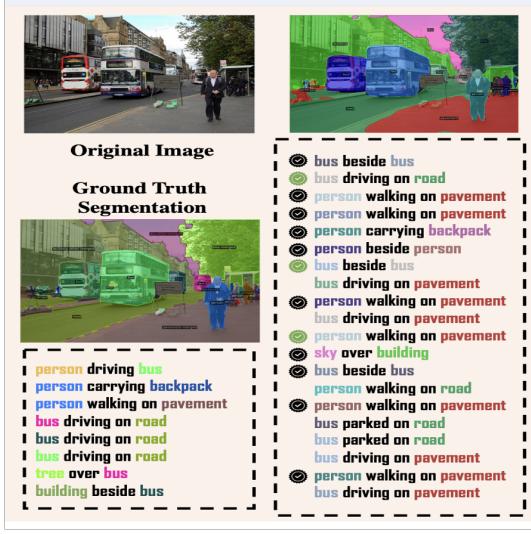


Figure 1. Panoptic Scene graph representation

2. Related Work

2.1. Advancements in Vision Language Models

In recent years there has been significant advancement in the field of vision language models with the inclusion of large language models leading to breakthrough performance in a variety of tasks. CLIP (Radford et al., 2021) used vision and language backbones to get embeddings and then fine tuned the embeddings in a way that would enable the model to match text-image pairs out of a large candidate batch. BLIP-2 (Li et al., 2023) takes a slightly different approach. It uses a lightweight querying transformer (Q-former) to bridge the image and language modalities. During pretraining, the Q-former is optimized on three different objectives: image-text matching, image text contrastive learning, and image grounded text generation. Consequently, BLIP-2 learns robust language grounded image embeddings, which it inputs to a Large Language Model (LLM) to generate text.

BLIP2 was selected for its advanced capabilities in bridging vision and language understanding. It excels in interpreting and contextualizing visual data within a linguistic framework, making it highly suitable for tasks requiring nuanced image-text associations.

2.2. Visual Abductive Reasoning

Sherlock (Hessel et al., 2022) is most similar in its task to visual reasoning datasets like VCR (Zellers et al., 2019), Visual Comet (Park et al., 2020), and Visual7W (Zhu et al., 2016). However, the Sherlock data set offers a unique set of challenges as it adopts a free viewing paradigm and isn't based on human centric examples. This format supports a wide range of topics through an open-ended annotation

approach. Sherlock also diversifies from human-centric annotations, allowing inferences to be grounded on any number of visual objects in an image, from central figures to background details. This approach fosters richer and softer notions of reasoning, expanding the scope of abductive reasoning in visual contexts beyond the constraints of previous datasets.

VAR still serves as a difficult benchmark for Vision Language Models. In the Sherlock dataset, current models significantly underperform compared to human benchmarks. This gap suggests that while models like CLIP show promise, they still struggle with the nuanced aspects of VAR that humans excel at, such as drawing complex inferences from subtle visual cues and contextualizing these cues within broader scenarios. The inherent complexity and subjectivity of abductive reasoning, combined with the challenge of understanding and integrating diverse visual and textual elements, contribute to this performance disparity.

2.3. Advancements in Fine-Grained Image Grounding Techniques for VAR

Initial advancements in fine-grained image grounding for Visual Abductive Reasoning (VAR) were significantly influenced by the introduction of Colorful Prompt Tuning (CPT) (Yao et al., 2022). This approach focused on enhancing VLMs by enriching the prompt engineering process, where color was used as a metaphorical element to signify the varying degrees of contextual information fed into the models. Colorful Prompt Tuning represented a shift towards a more nuanced integration of visual and textual data, aiming to improve the model's understanding of complex visual scenes.

Building upon the foundations laid by Colorful Prompt Tuning, the Regional Prompt Tuning (RGP) method (Zhang & Fernando, 2023) introduces a more sophisticated approach. RGP advances the concept of fine-grained image analysis by encoding regional visual hints and global contexts separately, allowing the model to capture detailed aspects critical for abductive reasoning. This method employs a Dual-Contrastive Loss, a key innovation that aligns visual features with factual descriptions and plausible hypotheses more effectively. Demonstrated through extensive experimentation on the Sherlock dataset, RGP shows a marked improvement over existing models, highlighting its effectiveness in handling VAR tasks.

2.4. Scene Graphs

A scene graph is a topological representation of a scene, the primary goal of which is to encode objects and their relationships. Moreover, the key challenge task is to detect/recognize the relationships between the objects. Traditional scene graph generation models (Johnson et al., 2015)

uses bounding boxes to detect objects followed by the prediction of their pairwise relationships. Recent years, more fine-grained scene graph generation techniques have been proposed. ORGC (Overlap Region and Geometrical Center) (Zhao et al., 2022) leverages overlap region features and geometric center information to enhance fine-grained relationship identification. And PSGFormer (Yang et al., 2022) proposes the task of Panoptic Scene Graph generation instead of traditional bounding box based scene graph generation, resulting in more comprehensive scene-graph results.

For our work we choose to use the PSGFormer (Yang et al., 2022), because it addresses the following problems which exist in traditional bounding-box based grounding techniques:

- **Coarse localization:** bounding boxes cannot reach pixel-level accuracy
- **Inability to ground comprehensively:** bounding boxes cannot ground backgrounds
- **Tendency to provide trivial information:** current datasets usually capture frivolous objects like head to form trivial relations like person-has-head, due to too much freedom given during bounding box annotation.
- **Duplicate groundings:** the same object could be grounded by multiple separate bounding boxes

3. Proposed Approach

3.1. Model

For our model we choose BLIP-2 (Li et al., 2023). BLIP-2 has shown strong performance on a variety of Vision Language tasks and is a state of the art model for several benchmarks. However, adapting BLIP-2 for VAR as a retrieval tasks posed a few challenges. Most prominently, BLIP-2 isn't a joint vision-language encoder only model like CLIP (Radford et al., 2021), so to adapt the model to the same pretraining task, we had to obtain the image and text embeddings separately from different parts of the model. Additionally, the BLIP-2 language model obtains text embeddings for language modelling so it maintains the sequence length dimension. To tackle this we used an adaptive average pooling layer to obtain an embedding for the entire text sequence, we then use a linear layer to project this to the same embedding dimension as the Q-former image embedding. The model architecture and our methodology is summarized in Figure 2.

3.2. Scene Graph Generation

To produce scene graphs, the PSGFormer (Yang et al., 2022) was employed. When presented with an image and a bounding box, our initial step involves crafting a scene graph that

encapsulates all relationships within the image. An example of the output of the PSGFormer is shown in Figure 3. The model responsible for scene graph generation outputs relationship triplets in the format of $(subject, predicate, object)$, where each subject/object aligns with a segmentation mask. Assessing each segmentation mask involves computing the overlap area $A_{overlap}$ between the mask and bounding box. Subsequently, we designate a scene graph triplet as pertinent if either the object or subject within the triplet satisfies either of the following conditions:

$$\frac{A_{overlap}}{A_{seg}} \geq 0.5 \quad \text{or} \quad \frac{A_{overlap}}{A_{bbox}} \geq 0.5$$

We exclusively generate relevant scene graph triplets within the target inference area (bounding box region) of the image. These triplets are then converted into text format, presented as a sequence of subject-predicate-object triples, and directly fed into the LLM encoder.

3.3. Training

3.3.1. ADAPTER BASED FINETUNING

To fine-tune BLIP-2 we decided to pursue an adapter based fine tuning approach. First introduced as PALs (Stickland & Murray, 2019) and recently popularized with LoRA (Hu et al., 2021). LoRA (Low-Rank Adaptation) is a parameter efficient fine tuning technique for large-scale pre-trained models. It introduces trainable low-rank matrices to modulate the existing weights in key layers, such as self-attention and feed-forward layers, of transformer models. During the training process, these low-rank matrices are the only parameters that are updated, while the original pre-trained weights remain frozen. This method allows for significant adaptations to new tasks with minimal changes to the overall model, thus preserving the strengths of the original pre-trained model while enabling effective customization for specific applications. For our training configuration we attach lora adapters to the BLIP-2 Q-former and LLM Encoder keeping the image encoder weights frozen.

3.3.2. INFONCE

In adapting the BLIP-2 model for the Sherlock dataset, our training task involves aligning complex visual scenes with corresponding textual inferences and scene graphs to enhance visual abductive reasoning capabilities. This task is ideal for the Sherlock dataset, which requires deep understanding beyond the explicit content of images in the form of discriminating between similar inferences. We employ the contrastive InfoNCE loss function (van den Oord et al., 2019), which is pivotal in strengthening the model's ability to identify nuanced, contextual correlations between images and text. The InfoNCE loss function, used in our adaptation for the Sherlock dataset, operates by contrasting pairs

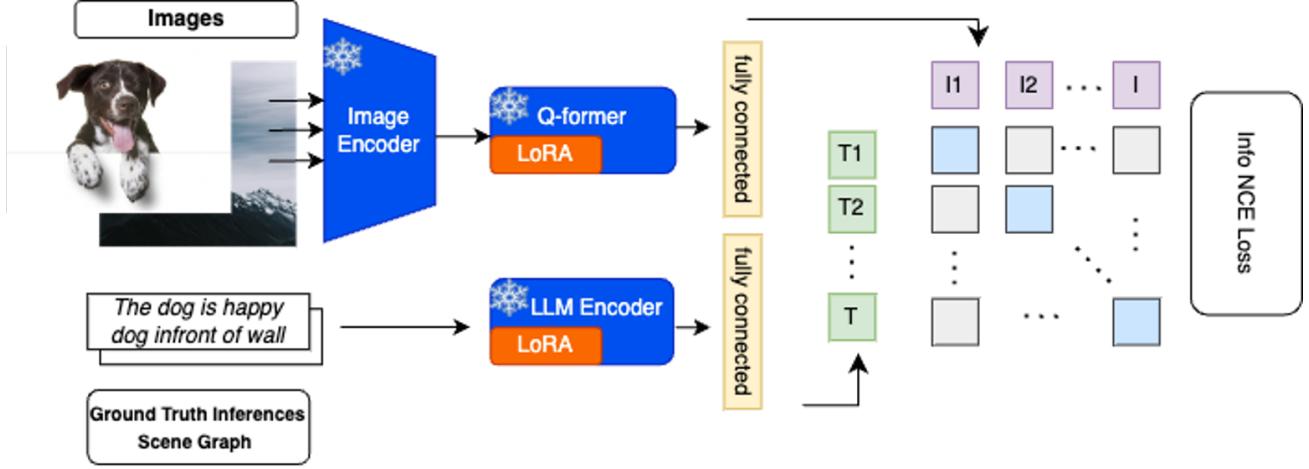


Figure 2. Adapting BLIP-2 for VAR Inference: our proposed approach: we freeze the weights of all components of the BLIP-2 model but fit the Q-former and LLM encoder with LoRA adapters. we input the scene graph information along with the ground truth caption, and use Infonce loss to promote the positive pairs and distance embeddings from negative pairs.

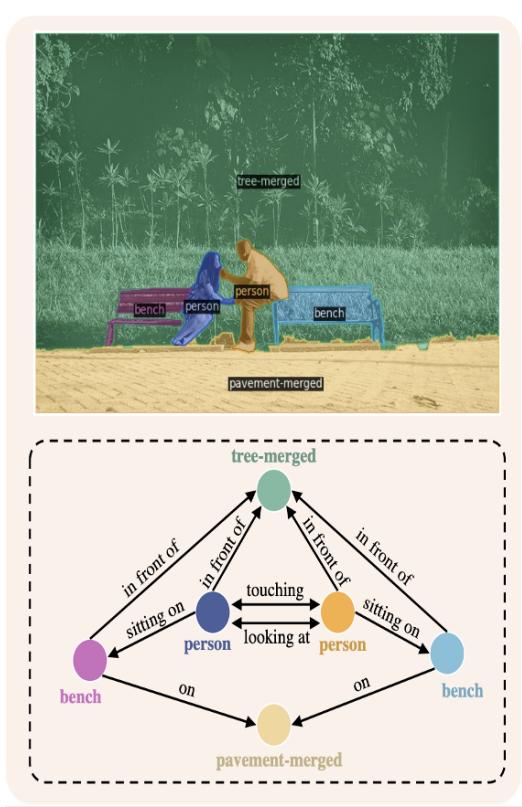


Figure 3. PSGFormer generated scene graph showing fine grained image grounding for the scene.

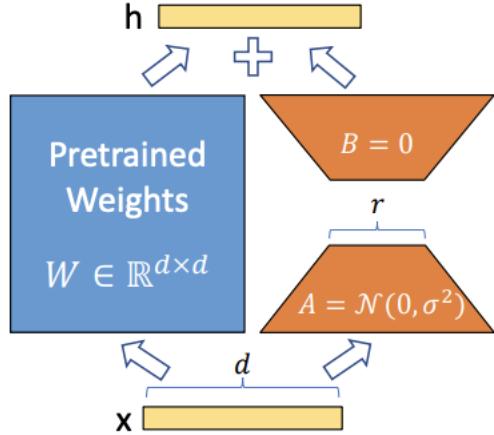


Figure 4. LoRA Adapters: This figure illustrates the integration of LoRA adapters in transformer models, demonstrating how they update key layers with minimal parameter changes while preserving original model weights.

of image-text data. In each training batch, the model predicts which text descriptions (consisting of the ground truth inference and scene graph) correctly correspond to which images. The loss function maximizes the similarity for correct pairings and minimizes it for incorrect ones.

This contrastive approach is highly effective in the Sherlock dataset context, as it forces the model to understand and align complex visual scenes with their nuanced textual inferences and fine grained scene graph information, crucial for abductive reasoning. This method not only enhances the model's accuracy in identifying correct pairings but also

improves its ability to discern subtler contextual details essential for abductive reasoning from visual data.

Given a set $X = \{x_1, \dots, x_N\}$ of N random samples comprising one positive sample from $p(x_{t+k}|c_t)$ and $N - 1$ negative samples from the 'proposal' distribution $p(x_{t+k})$, we optimize the image and text embeddings by minimizing the following contrastive loss:

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{e(x_{t+k}, c_t)}{\sum_{x_j \in X} e(x_j, c_t)} \right] \quad (1)$$

where $e(x, c)$ denotes the joint embedding function for image x and context c . Optimizing this loss will result in $e(x_{t+k}, c_t)$ approximating the density ratio, which is:

$$e(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})} \quad (2)$$

This density ratio reflects the alignment accuracy of the corresponding image-text pair, central to the abductive reasoning tasks within the Sherlock dataset. This approach ensures that the model not only captures surface-level details but also infers broader situational contexts, as well as fine grained details about the target inference region, a key requirement for effective abductive reasoning as demonstrated in our work.

3.4. Other Research Ideas: Abductive Reasoning as a Generative Task

Traditionally, visual abductive reasoning has been approached as a classification or retrieval task, with success measured by selecting or retrieving the correct inference from predetermined options. This approach, while effective for assessing specific reasoning skills, potentially constrains the model's ability to generate novel inferences. We propose a paradigm shift to view abductive reasoning as inherently generative, mirroring the creative human process of hypothesizing beyond visible data. This perspective aligns with the multifaceted nature of abductive reasoning, where multiple plausible inferences are valid. A generative model, informed by image clues, seeks to construct inferences that, while absent from the image, remain contextually grounded and diverse, thus expanding the model's inferential capacity.

4. Experimental Setup

4.1. Dataset and Input Modalities

The Sherlock corpus contains a total of 363K abductive commonsense inferences grounded in 81K Visual Genome and 22K Visual Commonsense Reasoning (VCR) images. For our experiments, we took a 20K subset of Sherlock dataset for training and 1K for validation and testing each.

Images in the dataset have an average of 3.5 observation

pairs, each training instance of data consists of the following components:

- An image i
- N bounding boxes $r = \{< x_{1i}, x_{2i}, y_{1i}, y_{2i} >\}$
- N clues c each corresponding to a literal description of r 's contents
- Inference f that is associated with i, r , and c

The model should take in an image i , a region r , and a candidate inference f , and the model's goal is to output a score s where s is proportional to the plausibility that f could be inferred from (i, r) . The clues are only used for training, not for the final task. We focus on the paradigm of retrieval of abductive inferences Give the model 1K candidate inferences, and take argmax of the inference that has the highest score (s). We evaluate the overall score of this task by taking the mean rank of the correct inference (lower means better) and the $P@1$.



Figure 5. An example from the sherlock dataset. The image has $N=3$ bounding boxes with associated clues c . There are three human annotated ground truth inferences f .

4.2. Multimodal Baseline Models

The original Sherlock paper used these baseline models: RN50x64-multitask and ViT/B-16.

RN50x64-multitask is the most performant model based on the original paper, and ViT/B-16 is the fastest and relatively light-weight model.

- RN50x64-multitask: RN50x64-multitask is one of the central models in our baseline investigation. It's a highly performant model, which was originally introduced in the

Sherlock research paper. This model is noteworthy for its versatility and robustness in handling a wide range of tasks, making it a valuable choice for various applications.

- ViT/B-16: ViT/B-16, on the other hand, stands out for its efficiency. It is known for being a fast and relatively lightweight model. This characteristic makes it an excellent candidate for scenarios where computational resources are limited or real-time processing is crucial.

The performance of baseline models are shown in Table 1. CLIP RN50x64 outperforms other models, but significant area for improvements exists.

	im → txt (↓)	txt → im (↓)	P@1(↑)
Random	495.4	495.4	0.1
Bbox Position	257.5	262.7	1.3
CLIP ViT-B/16	19.9	21.6	30.6
CLIP RN50x16	19.3	20.8	31.0
CLIP RN50x64	16.4	19.7	31.8

Table 1. Test results for baseline models for retrieval task. CLIP RN50x64 outperforms all models in all setups, but significant headroom exists.

4.3. Experimental Methodology

We use the opt variant of BLIP-2 which differs from the FlanT5 variant in that it uses a decoder only LLM rather than an encoder-decoder one. We load the model using 8 bit quantization for memory efficiency.

Model Configuration	
Model variant	BLIP-2
parameters	opt
Quantization	2.7 B
	8 bit

Table 2. Model configuration parameters:

For the LoRA configuration r controls the rank of the low rank matrix, it represents a tradeoff between flexibility and parameter efficiency. A smaller rank results in fewer parameters, making the model more efficient but potentially less flexible. The α parameter is a scaling factor applied to the LoRA matrices. This scaling allows for controlled adjustment of the impact these matrices have on the model's pre-trained weights, influencing the extent to which the original model is modified during adaptation.

Since we're conducting parameter efficient fine tuning we choose a low learning rate of 1e-5 and only train for 10 epochs. We choose the Adam optimizer with weight decay regularization set to 1e-5.

LoRA Configuration	
r	16
Alpha	32
' Dropout	0.05
Bias	None

Table 3. LoRA configuration parameters

Training Configuration	
Batch Size	64
Learning Rate	1e-5
Epochs	10
Optimizer	AdamW
Weight Decay	1e-2
Criterion	InfoNCE
Image Resolution	224 x 224

Table 4. Training configuration

5. Results & Discussion

5.1. Fine-grained Abductive Reasoning using scene graph

P@1 (Precision at 1):

Refer to Figure 6 Left, the Precision at 1 (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:

Refer to Figure 6 Right, 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.

Comparison to Baseline models and SoTA models:

From Table 5, we showed that our model outperforms the baseline models proposed in the original Sherlock paper both in terms of $P@1$ and mean rank. However, it does not outperform existing SoTA models which used region-prompt adapter tuning for VAR. It could be due to the lack of training resource we have during the model fine-tuning. And for future work, we can explore the performance of

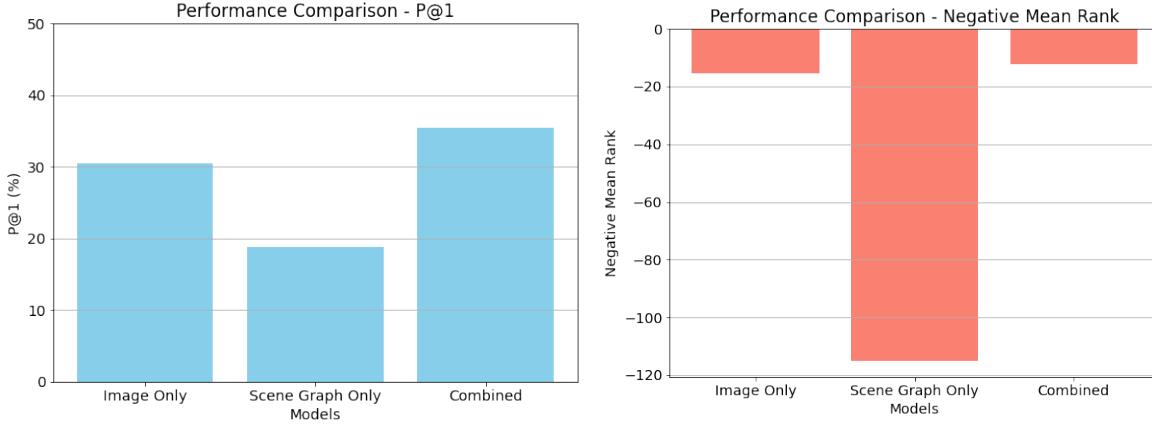


Figure 6. Left: the Precision at 1 ($P@1\uparrow$) scores for three distinct models—Image Only, Scene Graph Only, and Combined. Right: the negative mean rank (\uparrow) for three models. For both images, the task is: given image with bounding box, retrieve the best inference from 1000 candidate inferences.

our model and the model trained with Region-Prompted Adapter Tuning when given similar training resources.

Why scene graph improves the results:

Scene graphs capture intricate relationships between objects in an image, providing a structured representation of the visual scene. Integrating this information into Blip-2’s reasoning process enhances its contextual understanding. The model can leverage these relationships to infer more nuanced and accurate abductive reasoning by considering not just individual objects but their interconnections.

For example, in Figure 7 Middle, although the bounding box only involves the people, humans’ inference can be about anything that is inferred based on the bounding box. In this case the ground truth inference is inferring the truck is a food truck based on the fact that people are queuing in front of it. Without the scene graph, the image only model retrieved “people are standing in the shade” inference, and the ground truth inference “this is a food truck” was ranked 51. However, with the scene graph input explicitly stating the relationship “people at food truck”, the model is able to retrieve the ground truth inference easily.

Another example is shown in Figure 7 Right, with scene graph making the relationship “pizza in a box” explicit, the model was able to find the ground truth inference “this is made in pizzeria and not homemade”, rather than a generic inference that involves pizza.

However, scene graph itself is not able to provide enough information for the model to retrieve the correct inference most of the time. It is evident in Table 6 in which scene-graph-only model having a $P@1$ of 18.8% and a much lower negative mean rank of -114.95 . This is because image can often have details not captured by scene graphs which lead to abductive inferences. For example, in Figure 7 Left, the

image has the text “blender” on the machine, which leads to a human-inference of “the device is a blender”. However, this information cannot be captured by our scene graph generation model.

Therefore, the combined modalities of image and scene graph gives the best results for the VAR task.

5.2. Results for other research idea: Generative Task

We chose two state-of-the-art models for evaluation: Salesforce/blip2-opt-2.7b and LLaVA-v1.5. These models were selected based on their performance in related tasks and their availability for experimentation.

Salesforce/blip2-opt-2.7b and LLaVA-v1.5 were loaded as pre-trained models without further fine-tuning. This decision was made to evaluate their inherent generative abilities without introducing additional training biases.

Evaluation Metrics:

To quantify the generative performance, we employed the BLEU metric to quantify the generative performance of the models in the context of abductive reasoning, as we can see from table 6.

Additionally, we utilized a Likert scale for qualitative evaluation by human annotators to assess the credibility and contextual relevance of the generated abductions. This scale, ranging from 1 (strongly disagree) to 5 (strongly agree), asked whether the inference was related to the image or bounding box. The analysis was conducted internally within our team. Our findings indicated that although the context for LLaVA was more comprehensive, the context provided by BLIP-2 was more focused and relevant.

Results:

	im → txt (↓)	txt → im (↓)	P@1(↑)
CLIP ViT-B/16	19.9	21.6	30.6
CLIP RN50x16	19.3	20.8	31.0
CLIP RN50x64	16.4	19.7	31.8
RPA ViT-L-14 (Leaderboard Top)	10.1	12.6	40.3
BLIP-2 with SG	13.1	14.2	34.4

Table 5. Test results for BLIP-2 trained with scene graphs compared to baseline models and SOTA model.

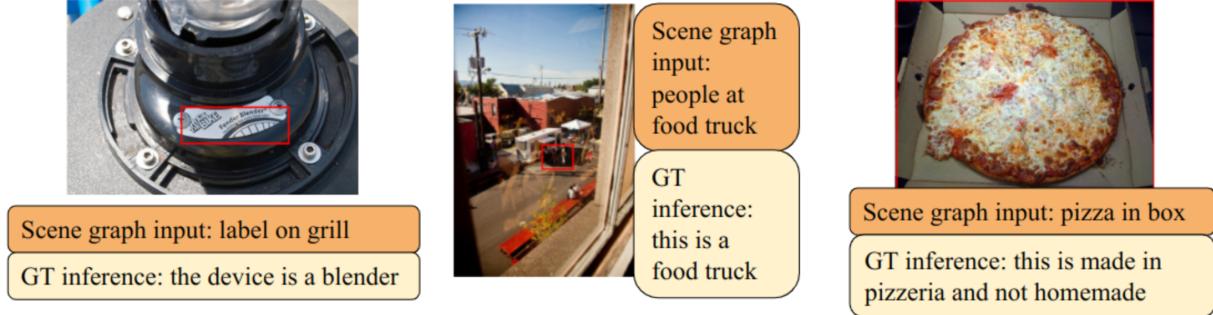


Figure 7. Left: example which image-only model retrieved correctly but not scene-graph-only model; Middle: example which scene-graph-only-model retrieved correctly but not image-only model; right: example which combined-model retrieved correctly but not uni-modality models

Evaluation Metrics	BLIP2	LLAVA
Avg BLEU Scores	0.83	0.45

Table 6. Evaluation Metrics for BLIP2 and LLAVA models

The output of LLaVA-v1.5 exhibited a higher level of contextual understanding, as evidenced by the images 10 and 12 (Appendix). In contrast, the output from Salesforce/blip2-opt-2.7b producing an inference that lacked the nuanced details present in the image , as evidenced by the images 11 and 13, yet it was more focused and relevant to the task.

6. Conclusion

In this study, we explored the role of scene graphs in enhancing Blip-2’s abductive reasoning capabilities for visual understanding tasks. Our findings reveal that incorporating scene graphs alongside image data significantly improves the Precision at 1 score, outperforming both image-only and scene graph-only models. The combined model showcases the strongest negative mean rank, signifying consistent and accurate prediction rankings. While scene graphs offer structured relationships and contextual understanding, they aren’t exhaustive in capturing all visual cues, leading to limitations when used in isolation. However, when synergistically employed with image data, they enrich Blip-2’s ability to infer nuanced and contextually grounded abductive reasoning.

Despite advancements, our model performance falls short of

the state-of-the-art models using region-prompted adapter tuning due to resource constraints during fine-tuning. Future research should explore these approaches with comparable resources to unveil their full potential. Nonetheless, our study highlights the promising role of scene graphs in advancing multimodal models’ reasoning capacities, fostering more contextually aligned visual abductive reasoning.

For our second idea, we assessed the generative abilities of two state-of-the-art models, Salesforce/blip2-opt-2.7b and LLaVA-v1.5, in the context of abductive reasoning. Through evaluation metrics such as the BLEU score and qualitative analysis via Likert scale ratings by human annotators, we measured the models’ performance. LLaVA-v1.5 demonstrated a higher level of contextual understanding, as evidenced by its generated inferences, whereas Salesforce/blip2-opt-2.7b exhibited limitations in capturing nuanced details within the images. These findings provide valuable insights into leveraging generative models for abductive reasoning tasks, paving the way for more sophisticated and context-aware AI systems.

References

- Chen, L., Chen, L., Ellison-Chen, T., and Xu, Z. On the cognition of visual question answering models and human intelligence: A comparative study, 2023.
- Hessel, J., Hwang, J. D., Park, J. S., Zellers, R., Bhagavatula, C., Rohrbach, A., Saenko, K., and Choi, Y. The abduction of sherlock holmes: A dataset for visual abductive reasoning, 2022.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.
- Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., and Fei-Fei, L. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
- Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., and Choi, Y. Visualcomet: Reasoning about the dynamic context of a still image, 2020.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.
- Stickland, A. C. and Murray, I. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning, 2019.
- van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding, 2019.
- Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., and Liu, Z. Panoptic scene graph generation, 2022.
- Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.-S., and Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models, 2022.
- Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning, 2019.
- Zhang, H. and Fernando, B. Fine-grained regional prompt tuning for visual abductive reasoning, 2023.
- Zhao, Y. Q., Jin, Z., Zhao, H. Y., Zhang, F., Tao, Z. W., Dou, C. F., Xu, X. H., and Liu, D. H. Fine-grained scene graph generation with overlap region and geometrical center. In *Computer Graphics Forum*, volume 41, pp. 359–370. Wiley Online Library, 2022.
- Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. Visual7w: Grounded question answering in images, 2016.

7. Appendix

7.1. Zero-shot results for VAR as generative task from BLIP-2 and LLaVA

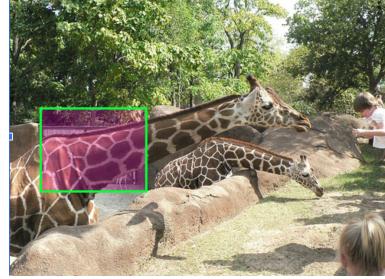


Figure 8. A zoo image chosen for evaluation



Figure 9. A restaurant scene chosen for evaluation

In the image, there are three giraffes in a zoo enclosure, with one of them being the main focus. A woman is standing near the giraffes, possibly feeding them or observing them. The giraffes are standing close to each other, and one of them is eating from a woman's hand. The scene suggests that the woman is interacting with the giraffes in a friendly and safe manner, possibly as part of a zoo activity or feeding program. The presence of the woman and the giraffes in the enclosure indicates that this is a controlled environment where visitors can have close encounters with the animals.

Figure 10. Inference generated by llava model for the giraffe image

| a giraffe is standing next to a wall

Figure 11. Inference generated by blip2 model for the giraffe image

From the image, we can infer that the two women are enjoying a meal together at an outdoor dining area, possibly at a restaurant or a café. They are seated at a dining table with chairs, and there are several cups, a bowl, and a fork visible on the table. The presence of multiple chairs and dining tables suggests that this outdoor space is designed for socializing and dining. The women seem to be having a pleasant time together, engaging in conversation or simply enjoying each other's company while sharing a meal.

Figure 12. Inference generated by llava model for the restaurant image

| a couple of women sitting on a porch

Figure 13. Inference generated by blip2 model for the restaurant image