One More Modality: Does Abstract Meaning Representation Benefit Visual Question Answering?

Abhidip Bhattacharyya
UMass Amherst
abhidipbhatt@umass.edu

Emma Markle
Amherst College
emarkle26@amherst.edu

Shira Wein
Amherst College
swein@amherst.edu

Abstract

Visual Question Answering (VQA) requires a vision-language model to reason over both visual and textual inputs to answer questions about images. In this work, we investigate whether incorporating explicit semantic information, in the form of Abstract Meaning Representation (AMR) graphs, can enhance model performance—particularly in low-resource settings where training data is limited. We augment two vision-language models, LXMERT and BLIP-2, with sentence- and document-level AMRs and evaluate their performance under both full and reduced training data conditions. Our findings show that in well-resourced settings, models (in particular the smaller LXMERT) are negatively impacted by incorporating AMR without specialized training. However, in low-resource settings, AMR proves beneficial: LXMERT achieves up to a 13.1% relative gain when using sentence-level AMRs. These results suggest that while adding AMR can inhibit VQA performance in some settings, AMR can serve as a useful semantic prior in a low-resource setting, especially for lower-capacity models trained on limited data.

1 Introduction

The task of visual question answering (VQA) (Antol et al., 2015; Malinowski and Fritz, 2014) challenges models to answer natural language questions about the content of an image. For example, given an image of a scene with sheep and the question "How many sheep are there?" the model must identify relevant visual elements and interpret the question to produce an accurate answer (see Figure 1).

Recent advances in multimodal learning have led to substantial gains in VQA performance (Huang et al., 2020; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Bai et al., 2023), powered by transformer-based vision-language models (VLMs) pretrained on large-scale image-text datasets. These models typically operate on raw textual and visual embeddings and lack integration of deeper, structured semantic representations. However, prior work in visual semantic parsing and caption generation (Hildebrandt et al., 2020; Lee et al., 2019; Yao et al., 2018; Chen et al., 2020a; Bhattacharyya et al., 2024) has demonstrated that incorporating formal semantic structures can enhance both the controllability and expressiveness of model outputs.

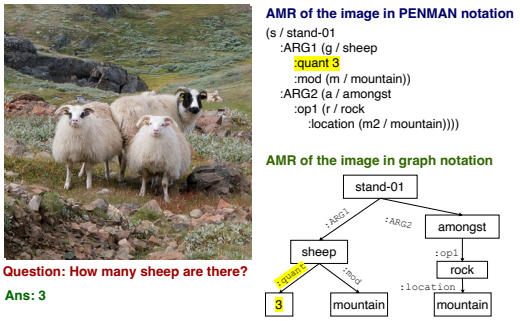


Figure 1: An example from the COCO-VQA dataset, where the AMR is generated from the longest image description available in the COCO captions. The AMR appears in PENMAN (text-based) notation (Kasper, 1989) and as a graph. Notably, the `:quant` node in the AMR encodes the correct answer to the question.

Given that VQA sits at the intersection of computer vision and language understanding, and promising prior work incorporating semantic representations into other vision-related tasks (Wein and Opitz, 2024), we explore the integration of a semantic representation as an additional modality for this task. Abstract Meaning Representation (AMR; Banarescu et al., 2013) is a rooted, directed graph-based representation of meaning that broadly captures "who does what to whom," where the nodes in the graph correspond to concepts in the sentence and edges denote the relationship between those concepts. We hypothesize that the AMR of an image description often encodes sufficient informa-

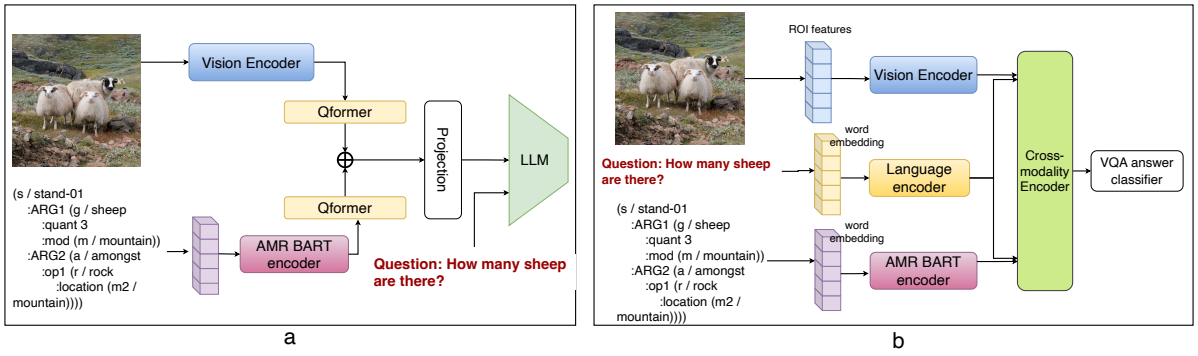


Figure 2: Updated model architectures for AMR-augmented vision-language models. (a) AMR integration in BLIP-2: a dedicated Q-former processes AMR embeddings, which are then combined with image features before being passed to the language model. (b) AMR augmentation in LXMERT: AMR embeddings from AMRBART are fed into a separate modality stream, which is later cross-attended by the language stream alongside visual features.

tion to answer associated questions. This motivates our central inquiry: **can integrating AMR into the VQA pipeline improve model performance, particularly under data or model capacity constraints**? To investigate this, we frame the following research questions:

RQ1 How does the presence of AMR in the training process impact VQA quality?

RQ2 How does model size impact the effectiveness of AMR in VQA?

RQ3 Can AMR compensate for limited image data in low-resource settings?

To investigate these questions, we augment VLMs with AMR graphs parsed from image descriptions in the COCO dataset (Lin et al., 2014). We compare two architectures: LXMERT (Tan and Bansal, 2019), a first-generation cross-modal transformer, and BLIP-2 (Li et al., 2023), a more advanced model that leverages frozen language models. To assess AMR's impact under varied resource conditions, we experiment with fine-tuning both sentence-level and document-level (docAMR) representations (Naseem et al., 2022), and analyze different configurations including frozen model components and varying contributions from visual versus semantic inputs. Our contributions include:

- comprehensive experimentation incorporating AMR graphs into a vision-language model for VQA, across various resource settings;
- a publicly available dataset[1] of both sentence and document-level AMRs of the image descriptions of COCO (Lin et al., 2014); and,
- analysis of the impact of incorporating AMR on the different kinds of questions comprised in the VQA task.

2 Methods & Experiments

In this section, we detail the data preparation procedures (Section 2.1), model configurations (Section 2.2), and evaluation metrics (Section 2.3) used in our experiments.

2.1 Data

In order to address our research questions and examine the effect of incorporating semantic information into a VQA model, we first collect AMR graphs of the image descriptions from the COCO dataset (Lin et al., 2014), as the VQA images are sourced from COCO. For fine-tuning, we use the VQA training set, which contains 443,757 question-answer pairs. We report our evaluation results on the VQA development set (214,354 question answer pairs).

We parse these AMRs both as (1) document-level AMR graphs (docAMR; Naseem et al., 2022), which contains intra-sentential coreferential information, as well as (2) standard sentence-level AMRs.

First, we generate the document-level AMR graphs. The COCO dataset (Lin et al., 2014) provides single-sentence captions for images, which often lack detailed information. To address this limitation, we generate multi-sentence image descriptions using the LLaVA-*v*1.5-7*b* model (Liu et al., 2023). Additionally, large vision-language models are known to hallucinate, particularly in longer generations, with errors frequently appearing toward the end of the output (Zhou et al., 2024;

[1] https://github.com/abhidipbhattacharyya/Image_AMR_VQA

Huang et al., 2024). To mitigate this, we remove any image-description pair where the length of the description exceeded the mean by more than three standard deviations. Then we parse the docAMRs of the remaining descriptions.

Next, for single-sentence AMRs, we select the longest available caption for each image from the COCO dataset and generate the corresponding AMR graph using a Transition-based AMR Parser (Drozdov et al., 2022).

For experiments under limited data conditions, we sample 10,000 examples from the full VQA training set. To ensure a representative subset, we adopt a cluster sampling approach. While the VQA dataset provides predefined question types (Antol et al., 2015), these categories are often coarse and fail to capture the semantic nuance of the questions. For example, some of the VQA-defined question types (per Antol et al. (2015)) such as *"what is this," "what is the,"* and simply *"what"* do not meaningfully differentiate between question intents. To obtain a more semantically informed grouping and perform a finer-grained question type differentiation, we apply k-means clustering to the question set. We encode each question as a dense vector[2] and then perform k-means clustering on these embeddings, setting the number of clusters to 100. From each cluster, we sampled 100 questions, resulting in a reduced training set of 10,000 examples with broad semantic coverage.

2.2 Models

We use the generated AMR graphs to fine-tune two vision-language models: BLIP-2 (Li et al., 2023) and LXMERT (Tan and Bansal, 2019). BLIP-2 integrates state-of-the-art vision transformers (ViTs) with large language models (LLMs) through a BERT-style encoder Qformer, while LXMERT combines Faster R-CNN-based (Ren et al., 2015b) image features with a BERT-base encoder (Devlin et al., 2019). This simpler design lacks the architectural depth and flexibility of BLIP-2, offering a useful point of comparison in evaluating the impact of AMR-based supervision.

To incorporate AMR information into the VLM pipeline for training and fine-tuning, we use AMRBART (Bai et al., 2022) as our AMR encoder. We apply a depth-first traversal of the AMR graphs (Bevilacqua et al., 2021; Hsu et al., 2023) to convert structural information into a sequence before feeding it into AMRBART. Moreover, following previous work (Bevilacqua et al., 2021; Hsu et al., 2023; Bai et al., 2022), we maintain an AMR-specific lexicon for tokenizing the flattened AMRs. The integration pipeline for BLIP-2 and LXMERT are illustrated in Figure 2.

In the case of BLIP-2, we introduce a separate Qformer dedicated to processing AMR representations. The outputs from the image Qformer and the AMR Qformer are averaged, then passed through a projection layer before being fed into the LLM. For our BLIP-2 experiments, we use an image resolution of 224×224.[3]

For LXMERT, we adopt the modality-specific architecture described in the original paper, where each modality—language, vision, and now AMR—is processed in a separate encoder stream. We introduce an additional AMR stream that encodes the AMR embeddings independently. As with the vision stream, the language stream attends to the AMR stream via cross-attention layers, enabling multimodal integration.[4] All training hyperparameters adhere to the original configuration, which are detailed in Appendices A and B.

2.3 Evaluation

For evaluation, we follow the work of Antol et al. (2015), which established the task of Visual Question Answering and an evaluation protocol, in selecting metrics. Accordingly, we use accuracy computed based on consensus matching between the predicted answer and the set of ground-truth human annotations. Accuracy is measured across four categories: *'yes/no'* for questions with binary answers, *'num'* for questions requiring a numeric response, and *'other'* for all remaining question types. An overall accuracy across all questions is reported under *aggregate* (agg).

We additionally perform a qualitative analysis

[2]Using the all-MiniLM-L12-v2 model from Sentence Transformers (Reimers and Gurevych, 2019)

[3]The original BLIP-2 implementation offers two vision backbone options: ViT-L/14 (Radford et al., 2021) and ViT-g/14 (Fang et al., 2023), as well as two choices for the language model: OPT (Zhang et al., 2022) and FlanT5-XL (Chung et al., 2022). In our setup, we adopt the ViT-g/14 vision encoder paired with the FlanT5-XL language model for all BLIP-2 experiments. For BLIP-2 we used the implementation given by salesforce. During fine-tuning, the parameters of both the ViT and the LLM are frozen. When this work is done the original BLIP-2 implementation does not include a training configuration file for the VQA task. Therefore, we reconstruct the YAML configuration using the details provided in the paper.

[4]For LXMERT, we use the implementation provided by the Hugging Face library. Visual features are extracted using Faster R-CNN, following the original LXMERT setup.

Model	agg	other	yes/no	num
BLIP-2(FT)	70.98	63.45	87.58	51.74
BLIP-2(pretrained) +docAMR	70.13	62.28	87.29	50.5
BLIP-2(pretrained) +AMR	70.27	62.52	87.33	50.57
LXMERT(FT)	84.64	78.92	96.19	73.11
LXMERT(FT) +docAMR	72.87	65	88.97	56.34
LXMERT(FT) +AMR	76.49	69.34	91.43	60.6
LXMERT(BASE) +docAMR	69.53	61.08	86.55	52.57
LXMERT(BASE) +AMR	71.19	63.3	87.46	54.29

Table 1: Performance of AMR-augmented BLIP-2 (Li et al., 2023) and AMR-augmented LXMERT (Tan and Bansal, 2019). All models are initialized from the pretrained BLIP-2 FlanT5-XL checkpoints and fine-tuned for five epochs on whole VQA training data. The LXMERT+AMR model is fine-tuned using both the Hugging Face (HF) VQA fine-tuned model and the pretrained model.

(Section 4) which explores how various components of the AMR graph contribute to individual instances of VQA results.

3 Results

We now present our experimental results addressing each of our three research questions.

3.1 RQ1: Presence of AMR on VQA Performance

To address RQ1, which concerns the impact of incorporating AMR into models for VQA, we conduct extensive experiments using both AMR and docAMR representations. We fine-tune the BLIP-2 model (Li et al., 2023) with each AMR variant and present the results in Table 1. Recall that we evaluate on accuracy as measured across four categories: *'yes/no'* for questions with binary answers, *'num'* for questions requiring a numeric response, *'other'* for all remaining question types, and an aggregate (*'agg'*) for all question types.

Our findings indicate that AMR-augmented versions of BLIP-2 do not outperform the original model. We hypothesize that this is due to the strong performance of the vanilla model, which benefits from large-scale image-text pretraining, whereas the AMR-augmented model may struggle to effectively integrate the new modality.

To verify this trend, we conduct a similar set of experiments using LXMERT (Tan and Bansal, 2019), with results shown in Table 1. The findings are consistent with those from BLIP-2, further suggesting that AMR integration poses challenges for VLM architectures out-of-the-box. However, a notable observation emerges: in LXMERT, the model with AMR has significant relative gain over model with docAMR. Therefore, we hypothesize that BLIP-2 is insensitive to AMR and tends to prioritize visual and textual signals over new modalities. In contrast, LXMERT, with its BERT-base architecture, shallower cross-modal fusion, and more limited context handling capacity, is better suited to shorter AMR inputs.

3.2 RQ2: Model Size on AMR Utility for VQA

To investigate RQ2, which examines the impact of model size on the effectiveness of AMR in VQA, we compare the relative performance of BLIP-2 and LXMERT, as shown in Table 1. In our experimental setup, the vanilla LXMERT model outperforms the vanilla BLIP-2 model. This finding contrasts with the original results reported in Li et al. (2023), where BLIP-2 achieves higher accuracy on the VQA task. This is likely due to a difference in model configuration, as the original training configuration for BLIP-2 VQA is not publicly available, leading us to develop a configuration file based on the information provided in Li et al. (2023).

Given this discrepancy, we focus on the relative performance of the AMR-augmented versions of these models, particularly in relation to model size. The AMR-augmented version of BLIP-2 has a significantly larger model size (approximately 4.25B parameters), compared to the AMR-augmented LXMERT (approximately 0.47B). For BLIP-2, the docAMR and AMR variants retain 98.89% and 98.90% of the performance of the vanilla model, respectively—indicating minimal degradation. In contrast, the AMR-augmented LXMERT models show greater performance drops. When initialized from the fine-tuned LXMERT, document-level and sentence-level AMR variants retain only 86.09% and 90.30% of vanilla performance, respectively. The gap is even wider when initialized from the pretrained (but not fine-tuned) LXMERT: 82.14% for document-level AMR and 84.10% for sentence-level AMR.

These results strengthen our previous hypothesis that larger models like BLIP-2 are more insensitive to the introduction of new modalities like AMR. In contrast, smaller models such as LXMERT are more sensitive to modality mismatches and require careful initialization and training to benefit from AMR integration. Related work has identified that various AMR integration techniques enable AMR graphs to be successfully leveraged for different

Model	agg	other	yes/no	num
BLIP-2	37.53	20.04	66.25	21.02
BLIP-2+docAMR	37.41	20.47	65.76	19.9
BLIP-2+docAMR+AMRBART	38.01	21.05	66.16	21.14
BLIP-2+AMR	37.27	20.17	66.17	18.75
BLIP-2+AMR+AMRBART	37.56	20.57	66.08	19.76
LXMERT	30.08	5.9	62.63	27.21
LXMERT+docAMR	24.27	0.58	63.76	0.24
LXMERT+docAMR+AMRBART	32.15	9.16	63.69	28.25
LXMERT+AMR	34.38	12.86	64.04	30.29
LXMERT+AMR+AMRBART	34.02	12.66	63.78	29.02

Table 2: Performance of VLMs with AMR in a low-resource setting (10k training samples). '+AMRBART' indicates that the AMRBART encoder is fine-tuned, as opposed to being kept frozen along with the ViT and LLM components. All models here are trained from scratch.

resource conditions, but that the linguistic information contained within an AMR graph is especially useful for compensating for limited pretraining data (Wein and Opitz, 2024).

3.3 RQ3: AMR Utility for VQA in Low-Resource Settings

As discussed in Sections 3.1 and 3.2, the integration of AMR does not yield substantial performance improvements—and even harms performance—in full-data settings, regardless of the model size. To investigate our third research question, which asks whether AMR can be beneficial in low-resource scenarios, we conduct additional experiments using a reduced training set of 10,000 samples. This reduced sample represents only 2.25% of the original training data. To ensure a balanced subset, we cluster the data by question type and sample evenly from each cluster (details of the cluster sampling can be found in subsection 2.1). We then train each model from scratch using the reduced dataset. This approach removes the benefit of pretraining on large-scale image-text corpora, thereby simulating a true low-resource scenario. The results of the experiments using this reduced training set are presented in Table 2. While the size of the training set is reduced, the size of the development set (which we use for evaluation) is kept unchanged.

In the low-resource setting, incorporating AMR leads to noticeable performance gains. For the BLIP-2 model, the docAMR configuration with a trainable AMRBART module outperforms the vanilla version. The BLIP-2 model performance when sentence-level AMR is incorporated into fine-tuning exhibits marginal improvements, varying depending on whether AMRBART is frozen or trainable. In contrast, for the LXMERT model, the docAMR setup with frozen AMRBART degrades performance notably, showing a 19.3% relative drop compared to the vanilla model. When AMR is trainable, AMR-augmented LXMERT models show relative gains of 6.8% with docAMR and 13.1% with AMR. This demonstrates that when AMR is integrated into the VLM pipeline with a trainable component, models can effectively associate meaning representation-based cues in vision-language (VL) tasks.

Model	agg	other	yes/no	num
BLIP-2+AMR	70.27	62.52	87.33	50.57
frozen image Qformer	66.48	57.3	85.72	45.85
30% image	68.81	60.91	86.25	48.56
50% image	69.95	62.03	87.17	50.4

Table 3: Performance of AMR-augmented BLIP-2 under different settings, varying image contribution (30% or 50% image) and Q-former configuration. All models were initialized from pretrained BLIP-2 FlanT5-XL checkpoints and fine-tuned for 5 epochs.

In Table 3, we further examine how controlling the contribution of image features affects performance in the BLIP-2 model with AMR. When the image Q-former is frozen and its contribution limited to 50%, performance drops by 5.4% compared to the vanilla BLIP-2. Unfreezing the image Q-former while reducing its contribution to 30% results in a 3.5% relative improvement over the frozen setup. Increasing the image contribution further, with a fully trainable image Q-former, yields an additional 1.6% gain. This strengthens our assumption that BLIP-2 prioritizes visual features over AMR features, as mentioned in subsection 3.1.

These results show that BLIP-2 models are

Question Type	AMR Role	AMR Role Function
how many	:quant	quantity or numeric value
how many people are	:quant	quantity or numeric value
how many people are in	:quant	quantity or numeric value
what number is	:quant	quantity or numeric value
what time	:time	temporal context; time of event or state
what is in the	:part-of	compositonal structure and part-whole relations
what is on the	:location	spatial or locational context
what room is	:location	spatial or locational context

Table 4: Semantic correspondence between VQA questions and AMR graph roles.

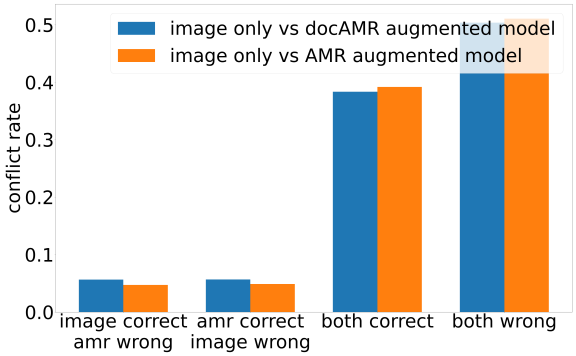
highly reliant on visual cues and struggle to effectively leverage semantic structures like AMR. This limitation may stem from the pretraining objectives of BLIP-2, which are optimized for learning from weak associations between image and text.

Overall, our experiments indicate that integrating AMR into VQA models poses challenges in full-resource settings. Larger models like BLIP-2 are relatively unaffected by the addition of AMR, while smaller models such as LXMERT may experience performance degradation in the full-data setting. The choice of AMR representation also matters: compared to the typical sentence-level AMR, BLIP-2 benefits more from incorporating docAMR due to its higher capacity and stronger fusion mechanisms, whereas LXMERT performs better with sentence-level AMR (relative to docAMR), likely due to its limited model size and simpler architecture.

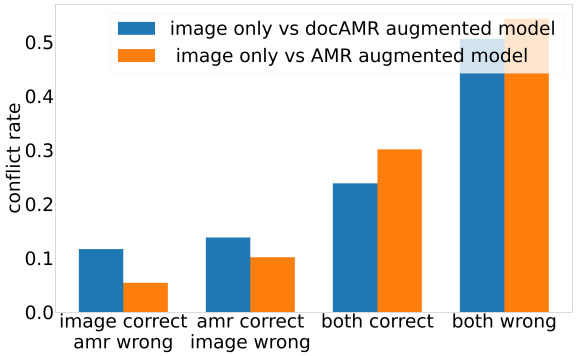
However, the utility of AMR for VQA becomes more apparent in the low-resource scenario. When training data is scarce, AMR provides consistent performance gains, acting as a valuable semantic prior that aids generalization. These findings highlight AMR's promise as an auxiliary modality for VQA, particularly under data constraints, and emphasize the importance of model-aware strategies for effectively integrating structured semantic representation.

4 Qualitative Analysis

To better understand model behavior in low-resource settings (as reported in Table 2), we visualize the conflict (disagreement) rate between predictions made by image-only models and their AMR-augmented counterparts (Figure 3). For LXMERT, the conflict rate differs noticeably between the AMR and docAMR variants, with the docAMR-augmented model exhibiting a higher rate of disagreement with the image-only baseline. In contrast, for BLIP-2, both AMR and docAMR variants show similar levels of conflict relative to the image-only model.



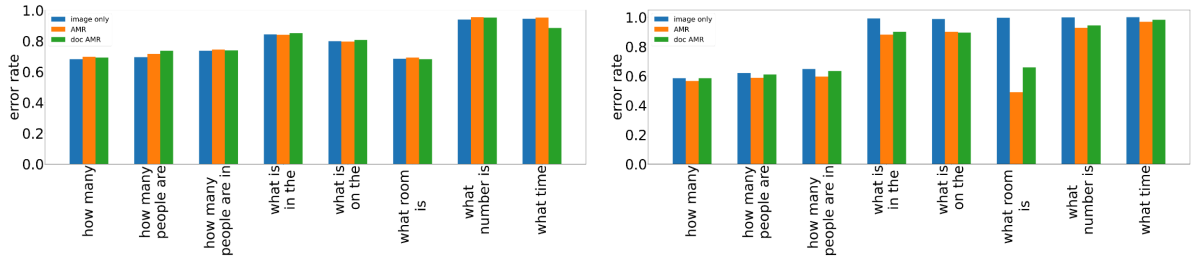
(a) BLIP-2



(b) LXMERT

Figure 3: Image only vs AMR conflict (disagreement) rate of BLIP-2 models (a) when compared with image only (vanilla) and AMR augmented variants. Similar experiments for LXMERT (b). All the models are trained from scratch with reduced data (model details found in Section 3).

Building on this, we set out to examine whether including AMR is more helpful in scenarios where the type of question aligns directly to a semantic role included in AMR, thus indicating that there is a specific piece of information included in the AMR that would correspond with the answer for that question type. We identify cases in which AMR's structured semantics align with specific VQA question types; to do so, we identify a subset of questions that exhibit explicit correspondences with AMR edge roles, as outlined in Table 4. For instance, questions beginning with *"How many"* often align with the `:quant` role, which encodes quantity-related information in AMR graphs. In the example shown in Figure 1, the `:quant` edge



(a) BLIP-2 (b) LXMERT

Figure 4: Error rates by question type (as listed in Table 4) for models trained in a low-resource setting (section 3). (a) shows results for BLIP-2, and (b) for LXMERT.

in the AMR graph contains the correct answer to the associated question. Similarly, questions like *"What is in the..."* may correspond to the `:part-of` relation, reflecting compositional structure.

As shown in Figure 4, error rates are similar across all question types for the three models. We observe that AMR-augmented models exhibit slightly lower error rates than their image-only counterparts for the question types *"What is in the..."* and *"What is on the..."* across both BLIP-2 and LXMERT. This suggests that AMR may help retain part-whole relationship information, likely due to the explicit use of the `:part-of` role in AMR graphs. We also note modest performance gains for the question *"What time..."* which corresponds to the `:time` role, particularly in the docAMR variant. Overall, the error rate reductions are more pronounced in LXMERT than in BLIP-2. These observations support our earlier findings: LXMERT, with its smaller model size and shallower cross-modal attention, is more sensitive to the structure of the AMR input and performs better with sentence-level AMR. Meanwhile, BLIP-2's deeper architecture and advanced fusion mechanisms make it relatively insensitive to the specific form of AMR used, maintaining stable behavior across both variants.

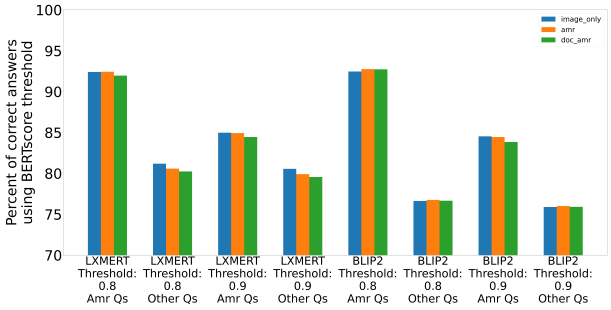
However, it is not always the case that we observe a lower error rate for AMR and docAMR on these questions identified as having explicit AMR roles. For example, while the AMR model achieves the lowest error rate on "how many" questions in LXMERT, it performs worse than the image-only baseline for the same question type in BLIP-2.

In addition to raw error rate, we assess the quality these answers (and thus the models with and without AMR) via BERTScore (Zhang et al., 2019) and BLEU (Papineni et al., 2002), comparing predicted answers against ground truth responses. A prediction is considered correct if its similarity score exceeds a predefined threshold. We then calculate the percentage of correct answers for each model variant under the low-resource setting.

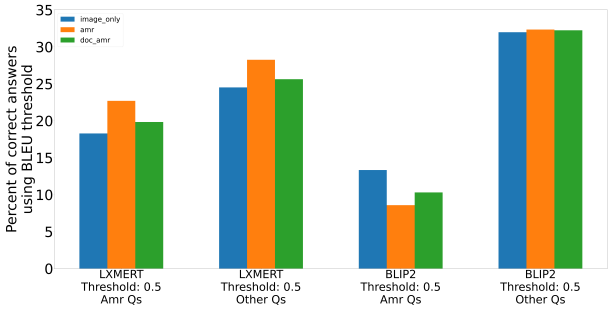
For BLIP-2, consistent with earlier findings, the benefits of AMR augmentation are minimal and largely unnoticeable under both BERTScore and BLEU-based thresholding. In contrast, LXMERT shows noticeable improvements (see Table 4) when evaluated using BLEU. However, improvements are less apparent with BERTScore, likely because it assigns high similarity scores to words that are close in the embedding space, yet semantically incorrect—such as numerical values (e.g., three vs. four) or binary opposites like yes and no—thus reducing its discriminative efficacy. Conversely, BLEU's n-gram-based matching penalizes such lexical deviations more strongly, which can be beneficial in filtering out answers that appear close via vector embeddings but are semantically incorrect. These trends are illustrated in Figure 5.

While performance gains are not uniform across all question types, our analysis indicates that incorporating AMR can lead to targeted improvements, particularly when the semantics of a question aligns closely with explicit AMR edge roles. These findings underscore the potential of AMR to preserve and leverage specific semantic information, especially in low-resource settings where visual context alone may be insufficient.[5]

[5]In a pilot study, we also prompted a language-only model (LLaMA-3-8B Instruct) with image AMRs in an in-context format. We provided the model with examples of several QA pairs along with their corresponding AMRs, followed by the actual question and the AMR representation of the image. However, the evaluation results were poor because LLaMA did not follow the expected format, and there was a noticeable vocabulary shift.



(a) BERTScore



(b) BLEU

Figure 5: Semantic answer evaluation under low-resource settings using BLEU and BERTScore. (a) BLIP-2 results show limited gains from AMR augmentation. (b) LXMERT exhibits noticeable improvement with AMR-aligned questions in BLEU scores, especially under stricter thresholds, while BERTScore remains less sensitive due to high similarity scores for semantically close but incorrect answers.

5 Related Work

VQA (Antol et al., 2015; Malinowski and Fritz, 2014) has gained significant research momentum over the past decade. Early neural approaches primarily adopted encoder-decoder architectures (Malinowski et al., 2015; Ren et al., 2015a; Noh et al., 2016), combining CNN-based image encoders with RNN-based question decoders. Inspired by advances in attention-based machine translation (Bahdanau et al., 2015), prior work has begun integrating visual attention mechanisms to improve question grounding (Anderson et al., 2018; Kazemi and Elqursh, 2017; Sun and Fu, 2019; Jang et al., 2017; Xu and Saenko, 2016; Zhu et al., 2017, 2015).

More recently, the rise of large-scale image-text datasets and the success of transformer-based VLM pretraining approaches have significantly boosted performance in vision-language tasks (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2019; Su et al., 2020; Li et al., 2021, 2020; Chen et al., 2020b). These developments, along with breakthroughs in large language models (Raffel et al., 2020; Lewis et al., 2020; Wei et al., 2022; Brown et al., 2020; Touvron et al., 2023) and VLMs (Radford et al., 2021; Jia et al., 2021), have led to the emergence of multi-stage, modular architectures—often referred to as 'socratic' models (Zeng et al., 2023). In such architectures, a vision model and a language model are connected via intermediate projection or reasoning modules that enable interaction while keeping both components largely frozen. Recent works that adopt this paradigm (Mokady et al., 2021; Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Alayrac et al., 2022; Bai et al., 2023, 2025b,a) typically use frozen vision encoders and LLMs, bridging them with lightweight interface modules designed to align their representational spaces.

However, current VLMs largely ignore explicit high-level semantic representations. Prior work has explored scene graphs (Xu et al., 2017; Tang et al., 2020, 2018) and their utility in downstream vision-language tasks (Hildebrandt et al., 2020; Lee et al., 2019; Ghosh et al., 2019; Yao et al., 2018; Chen et al., 2020a). More recent research highlights the value of structured semantic representations—such as semantic role labeling (SRL; Palmer et al., 2005) and AMR (Banarescu et al., 2013)—in improving performance on downstream multimodal tasks (Choi et al., 2022a,b; Abdelsalam et al., 2022; Bhattacharyya et al., 2022, 2024). Additionally, as is shown in this work, AMR has been shown to be particularly useful for engineering tasks in low-resource settings (Wein and Opitz, 2024).

The integration of semantic representations into modern pretrained VLMs remains underexplored; in this work, we explore the impact of incorporating high-level semantic representations, specifically AMR, into VLMs for the targeted task of VQA–with a focus on leveraging AMR in low-resource VQA settings.

6 Conclusion

In this work, we explore the utility of incorporating semantic information, specifically Abstract Meaning Representation, as an additional modality in the task of VQA. Our findings indicate that current VLMs do not consistently benefit from AMR augmentation in standard training settings. Experimenting on both a first-generation model (LXMERT; Tan and Bansal, 2019) and a more advanced model (BLIP-2; Li et al., 2023), we observe that the absence of AMR-aware pretraining makes it challenging for these models to

effectively integrate semantic structures as an additional modality. However, in our low-resource experiments using only 2.25% of the original training data, AMR integration consistently improves performance across both model architectures. This suggests that AMR can serve as an effective inductive bias, particularly when training data is limited, to compensate for a lack of visual data.

One bottleneck in incorporating AMR into the VQA pipeline is the lack of readily available AMRs for images. Developing an effective system for parsing images directly into AMR-like semantic representations (Abdelsalam et al., 2022) is a promising direction for future work. We generate silver-standard sentence- and document-level AMRs for the VQA dataset to support reproducibility and future research on AMR for VQA.

Limitations

Our approach relies on generating AMR from textual descriptions of images, which introduces several sources of potential error. We automatically parse the docAMR graphs, after generating multi-sentence image descriptions. Automatically parsing AMR graphs introduces noise and as a result, any limitations, hallucinations, or inaccuracies in the generated text, along with any parsing errors from the AMR system, propagate through our pipeline and may affect downstream VQA performance. For sentence-level AMR, although the input sentences are human-annotated captions from the COCO dataset, they are also automatically parsed using the IBM AMR parser, which may still introduce structural inaccuracies. These compounded errors in semantic parsing and text generation may limit the effectiveness of AMR as an additional modality, particularly (as we observe) in high-capacity models that already rely heavily on image-text alignment. Additionally, automatically parsing a large number of AMR graphs at test-time can be quite time consuming, leading to scalability issues.

References

Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat Bhatt, Vladimir Pavlovic, and Afsaneh Fazly. 2022. Visual semantic parsing: From images to Abstract Meaning Representation. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 282–300, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Lichen Bai, Zixuan Xiong, Hai Lin, Guangwei Xu, Xiangjin Xie, Ruijie Guo, Zhanhui Kang, Hai-Tao Zheng, and Hong-Gee Kim. 2025a. Frozen language models are gradient coherence rectifiers in vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1817–1825.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to rule them both: Symmetric AMR semantic parsing and generation without a complex pipeline. In *Proceedings of AAAI*.

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. 2022. Aligning images and text with semantic role labels for fine-grained cross-modal understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4944–4954, Marseille, France. European Language Resources Association.

Abhidip Bhattacharyya, Martha Palmer, and Christoffer Heckman. 2024. ReCAP: Semantic role enhanced caption generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13633–13649, Torino, Italia. ELRA and ICCL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020a. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9962–9971.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*.

Woo Suk Choi, Yu-Jung Heo, Dharani Punithan, and Byoung-Tak Zhang. 2022a. Scene graph parsing via Abstract Meaning Representation in pre-trained language models. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 30–35, Seattle, Washington. Association for Computational Linguistics.

Woo Suk Choi, Yu-Jung Heo, and Byoung-Tak Zhang. 2022b. Sgram: Improving scene graph parsing via abstract meaning representation. *Preprint*, arXiv:2210.08675.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Drozdov, Jiawei Zhou, Radu Florian, Andrew McCallum, Tahira Naseem, Yoon Kim, and Ramón Astudillo. 2022. Inducing and using alignments for transition-based AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1086–1098, Seattle, United States. Association for Computational Linguistics.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369.

Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. 2019. Generating natural language explanations for visual question answering using scene graphs and visual attention. *Preprint*, arXiv:1902.05715.

Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. 2020. Scene graph reasoning for visual question answering. *Preprint*, arXiv:2007.01072.

I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. 2023. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada. Association for Computational Linguistics.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *Proceedings*

of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7166–7176, Online. Association for Computational Linguistics.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*.

Robert T. Kasper. 1989. A flexible interface for linking applications to Penman's sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *Preprint*, arXiv:1704.03162.

Soohyeong Lee, Ju-Whan Kim, Youngmin Oh, and Joo Hyuk Jeon. 2019. Visual question answering over scene graph. In *2019 First International Conference on Graph Computing (GC)*, pages 45–50.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. Semvlp: Vision-language pre-training by aligning semantics at multiple levels. *CoRR*, abs/2103.07829.

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 *International conference on machine learning*, pages 19730–19742. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *ECCV 2020*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 1682–1690, Cambridge, MA, USA. MIT Press.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 1–9, USA. IEEE Computer Society.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Tahira Naseem, Austin Blodgett, Sadhana Kumaravel, Tim O'Gorman, Young-Suk Lee, Jeffrey Flanigan, Ramón Astudillo, Radu Florian, Salim Roukos, and Nathan Schneider. 2022. DocAMR: Multi-sentence AMR representation and evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3496–3505, Seattle, United States. Association for Computational Linguistics.

Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 30–38.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015a. Exploring models and data for image question answering. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA. MIT Press.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster r-cnn: towards real-time object detection with region proposal networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 91–99, Cambridge, MA, USA. MIT Press.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Qiang Sun and Yanwei Fu. 2019. Stacked self-attention networks for visual question answering. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 207–211, New York, NY, USA. Association for Computing Machinery.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2018. Learning to compose dynamic tree structures for visual contexts. *Preprint*, arXiv:1812.01880.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shira Wein and Juri Opitz. 2024. A survey of AMR applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. *CoRR*, abs/1809.07041.

Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. *Preprint*, arXiv:2310.00754.

Chen Zhu, Yanpeng Zhao, Shuaiyi Huang, Kewei Tu, and Yi Ma. 2017. Structured attentions for visual question answering. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1300–1309.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2015. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

A BLIP2 Parameters

[$^{E}_{M}$ title and text below should be BLIP-2, right? Not BLIP2?] Hyperparameters used for our BLIP2 experiments are summarized in Table 5.

param	value
vision model	ViT-g/14 (Fang et al., 2023)
language model	FlanT5 (Chung et al., 2022)
image size	224×224
init lr	$1e-5$
weight_decay	0.05
lr_sched	"linear_warmup_cosine_lr"
freeze_vit	True
prompt	"Question: {} Short answer:"

Table 5: Hyperparameter used for our BLIP2 [$^{E}_{M}$ once again BLIP-2, not BLIP2 for consistency] experiments.

B LXMERT Parameters

For LXMERT, we use the implementation provided by the Hugging Face library. Hyperparameters used for our LXMERT experiments are summarized in Table 6.

param	value
x_layers	5
l_layers	9
r_layers	5
hidden_size	768
learning_rate	$5e-5$
weight_decay	0.05
lr_sched	"linear_warmup_cosine_lr"
adam_epsilon	$1e-8$
visual_feat_dim	2048

Table 6: Hyperparameters used for our LXMERT experiments.

C LLaVA setup

The default image descriptions provided in the COCO dataset (Lin et al., 2014) are typically single-sentence captions, which often omit important visual details. To address this limitation, we generate multi-sentence descriptions using the LLaVA-v1.5-7B model (Liu et al., 2023). Hyperparameters used for generation are detailed in Table 7.

However, large vision-language models are known to hallucinate, especially in longer generations, with errors often occurring toward the end of the output (Zhou et al., 2024; Huang et al., 2024). To reduce the impact of such hallucinations, we filter out any image-description pairs where the description length exceeds the mean by more than three standard deviations.

param	value
models	llava-v1.5-7b
prompt	"Please provide a descriptive caption of the image in no more than 4 to 5 sentences. Include details such as the number of objects, their positions, relative placement, colors, and other attributes. Do not mention objects or concepts that are not present in the image. Avoid starting with phrases like 'The image depicts', 'The image features' or 'The image shows.' Keep the caption under 200 words."
sampling temperature	0.2
max-new-tokens	512

Table 7: LLaVA model hyperparameters for generating image description for document level AMR.