REPHRASE, AUGMENT, REASON: VISUAL GROUNDING OF QUESTIONS FOR VISION-LANGUAGE MODELS

Archiki Prasad **Elias Stengel-Eskin** **Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{archiki, esteng, mbansal}@cs.unc.edu

ABSTRACT

An increasing number of vision-language tasks can be handled with little to no training, i.e., in a zero and few-shot manner, by marrying large language models (LLMs) to vision encoders, resulting in large vision-language models (LVLMs). While this has huge upsides, such as not requiring training data or custom architectures, how an input is presented to an LVLM can have a major impact on zero-shot model performance. In particular, inputs phrased in an *underspecified* way can result in incorrect answers due to factors like missing visual information, complex implicit reasoning, or linguistic ambiguity. Therefore, adding visually-grounded information to the input as a preemptive clarification should improve model performance by reducing underspecification, e.g., by localizing objects and disambiguating references. Similarly, in the VQA setting, changing the way questions are framed can make them easier for models to answer. To this end, we present **Rep**hrase, **A**ugment and **Re**ason (REPARE), a gradient-free framework that extracts salient details about the image using the underlying LVLM as a captioner and reasoner, in order to propose modifications to the original question. We then use the LVLM's confidence over a generated answer as an unsupervised scoring function to select the rephrased question most likely to improve zero-shot performance. Focusing on three visual question answering tasks, we show that REPARE can result in a 3.85% (absolute) increase in zero-shot accuracy on VQAv2, 6.41%, and 7.94% points increase on A-OKVQA, and VizWiz respectively. Additionally, we find that using gold answers for oracle question candidate selection achieves a substantial gain in VQA accuracy by up to 14.41%. Through extensive analysis, we demonstrate that outputs from REPARE increase syntactic complexity, and effectively utilize vision-language interaction and the frozen LLM. [1]

1 INTRODUCTION AND MOTIVATION

Recent advancements in foundational vision-language (VL) models such as GPT-4 (OpenAI, 2023), BLIP-2 (Li et al., 2023), and Flamingo (Alayrac et al., 2022) have enabled tremendous strides in visual understanding tasks (Gan et al., 2022; Zhang et al., 2023a; Yin et al., 2023). Similar to large language models (LLMs) in the text domain (Ouyang et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023, *inter alia*), these large vision-language models (LVLMs) can be guided through well-designed input prompts to perform tasks without fine-tuning, i.e., in a zero- and few-shot fashion. This is a powerful capability, allowing models to be applied to vision-language tasks without access to large annotated training datasets. In this setting, the prompt's phrasing becomes crucial to model performance (Webson & Pavlick, 2021; Mishra et al., 2022; Prasad et al., 2023). Further contributing to the challenge of zero-shot tasks is *underspecification*, a common phenomenon in various VL tasks. In this work, we use visual question answering (VQA) as a representative VL task and seek to improve zero-shot model performance by addressing underspecification. In VQA, underspecified questions might provide inadequate information for an interlocutor to understand their intended meanings and answer them correctly (Pezzelle, 2023; Zhu et al., 2023a; Hu et al., 2022).

[1] Our code is puplicly available: https://github.com/archiki/RepARe

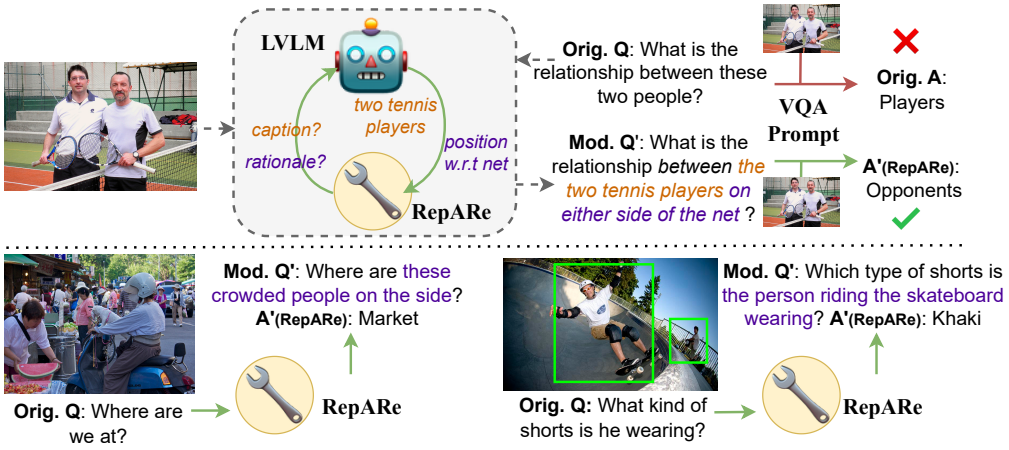

Figure 1: **Top**: The original question (in A-OKVQA) lacks information about implicit reasoning, leading to an incorrect answer. REPARE interacts with the LVLM to extract attributes like "tennis players" and "position w.r.t net" that are key to answering the question correctly. Adding these modifiers to the question elicits the correct response from LVLM. **Bottom**: Underspecified questions from A-OKVQA (left) and VQAv2 (right) datasets along with REPARE outputs.

Underspecification in VL tasks like VQA can manifest in several ways, leading to incorrect model predictions. Firstly, language lacking in visual details (i.e., questions *underspecified w.r.t. image*) can make it harder for models to align text and visual features (Pezzelle, 2023). For example, in Fig. 1 (bottom-left), *"we"* is not grounded to the image. Furthermore, abstract questions often require complex reasoning or external world knowledge that may not be present in the model or at least may be hard to access; in other words, the question is *underspecified w.r.t the world* (Marino et al., 2019; Schwenk et al., 2022). For instance, in Fig. 1 (top), the spatial arrangement of players on opposite sides suggests they are opponents. Explicitly referencing *"the net"* could help the model access this commonsense knowledge. While LVLMs might still rank *"opponent"* highly in their predictions for the original question, the rephrased question more clearly specifies the intent of the inquiry, leading to *"opponent"* as the generated response. Finally, some questions are inherently ambiguous, with multiple valid answers. Even if the model is capable of generating all possible responses, it is not clear which one is intended (Bhattacharya et al., 2019; Stengel-Eskin et al., 2023), i.e., the question is *underspecified w.r.t. intended meaning*. For example in Fig. 1 (bottom-right) there are two men, which results in multiple possible referents for *"he"*. Building upon prior research in textual question reframing (Dong et al., 2017; Majumder et al., 2021; Pyatkin et al., 2023), we hypothesize that making some of the details needed to answer the question more explicit could improve model performance.

There are several paths to addressing the challenges posed by underspecification. One approach involves additional VL pretraining to better align underspecified text to images as well as to enhance the LVLM's internal world model, enabling it to decifer underspecified questions in human-like ways. However, scaling up VL pretraining can be prohibitively expensive (Alayrac et al., 2022; Driess et al., 2023). Note that, in addition to the expense of finetuning, it could be that underspecification in the *training* data leads to continued subpar performance on underspecified questions. Another option is acquiring additional data or information from the user, such as clarifications. This strategy is infeasible for most standard VL benchmarks as they are static datasets (Kiela et al., 2021; Sheng et al., 2021). Furthermore, clarification interactions with users are time-consuming and costly. Thus, our method preemptively incorporates clarifications to reduce ambiguity, emphasize relevant visual details, and suggest reasoning steps, thereby, automatically improving the LVLM's VQA performance without the need for human intervention. Moreover, using preemptive clarifications could also hold value in VL dialogue systems, where users prefer concise interactions but often pose vague questions. This approach has several advantages: (i) it allows for a flexible, gradient-free framework to improve the performance of existing LVLMs without the need for additional pretraining or manual annotations; (ii) our text-based edits are human-readable i.e., we can verify that added details are relevant and consistent with the question's intent; and (iii) crucially, our method harnesses the *asymmetric strength* of most existing LVLMs, whose LLM components typically have far more ca-

pacity and pre-training data than the vision component,[2] and which often have strong reasoning and planning abilities on multimodal data (Wei et al., 2022b; Brohan et al., 2023; Guo et al., 2023). In other words, by preemptively rephrasing questions, we can align more closely with the strengths of existing LVLMs, making rich visual information from the image easier to access.

In Fig. 1 (top), we illustrate at a high level how rephrasing and modifying questions based on the image improves model predictions. Note that our method *does not* have any access to the gold answer, using model confidence to select a question. While the original question elicits a generic response, pinpointing the "*tennis players*" and emphasizing their positions "*relative to the net*" helps the model answer correctly. These modifications are obtained via self-interaction with the LVLM to get more information about the entities in the question as well as other salient objects from model-generated rationales and captions. To this end, we introduce **Rep**hrase, **A**ugment and **Re**ason (REPARE), a gradient-free, instance-level language adaptation framework to address underspecification. Broadly, REPARE consists of two stages: question rephrasing and augmentation, followed by question selection. First, we identify salient entities from the question and generate rationales as well as captions. These features help incorporate visually grounded information into the question. Conditioned on this information, we sample n modified question candidates including the original question. In the next stage, we utilize a confidence-based selection function to choose the most promising candidate, assuming that questions leading to higher-confidence answers are easier for the model to answer, and thus more likely to be correct. The overall pipeline is illustrated in Fig. 2.

Empirically, we show that REPARE improves zero-shot VQA performance by up to 3.85%, 6.41%, and 7.94% on the VQAv2 (Goyal et al., 2017), A-OKVQA (Schwenk et al., 2022), and VizWiz (Gurari et al., 2018) datasets, respectively using LVLMs including BLIP-2 (Li et al., 2023), MiniGPT-4 (Zhu et al., 2023b), and LLaVA-1.5 (Liu et al., 2023a) models in Sec. 4. Note that all percentages we report in this paper are *absolute* improvements. We further demonstrate the capabilities of REPARE in an oracle setting, establishing an upper-bound performance increase of up to 9.84%, 14.41%, and 20.09% on VQAv2, A-OKVQA, and VizWiz tasks, respectively. We extensively evaluate our design choices in Sec. 4.1 and quantitatively show the importance of incorporating visual information to address underspecification, as done in REPARE, compared to paraphrasing in Sec. 4.2. We analyze REPARE's outputs using linguistically-informed metrics like average dependency distance (Gibson et al., 2000) and idea density (Boschi et al., 2017). This reveals that the resulting questions are indeed less underspecified, i.e., more complex (see Sec. 4.3). Finally, in Sec. 4.4, we verify that questions from REPARE make better use of existing LVLMs by leveraging the strength of the LLM while still benefitting from the image. In summary, our contributions include:

- We propose REPARE, a novel zero-shot pipeline that interacts with LVLMs to modify underspecified questions by extracting and fusing information from keywords, rationales, and captions. This grounds questions in the image and commonsense knowledge while also making them less ambiguous, preemptively clarifying them to address underspecification without any human feedback.
- We empirically demonstrate that REPARE boosts zero-shot performance on three standard VQA benchmarks for a collection of LVLMs varying in model architecture, size and VL pretraining by up to 7.94%. Our oracle results suggest that we can obtain as high as 20.09% increase in zero-shot VQA accuracy *solely* by modifying the question.
- Extensive analysis shows that REPARE enhances question complexity via semantic modifications, outperforms paraphrasing, and harnesses LVLM's strengths with simple yet effective modules.

2 RELATED WORK

Large Vision-Language Models. Significant strides have been made in jointly processing language and images, especially in visual question answering. VQA (Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019; Johnson et al., 2017) has become a benchmark task for VL models. Recent methods address VQA as a zero- and few-shot learning task. These approaches can be categorized into two groups: (i) those relying on continuous image representations (Alayrac et al., 2022; Tsimpoukelli et al., 2021; Zhu et al., 2023b; Liu et al., 2023b; Li et al., 2023, *inter alia*); and (ii) those extracting linguistic information such as captions from images (e.g., Yang et al., 2022;

[2] E.g., BLIP-$2_{\text{Flan T5 xl}}$ (Li et al., 2023) consists of a ViT vision encoder, a Q-former fusion module, and a Flan-T5 LLM with 1B, 0.11B, 3B and parameters respectively, i.e., the LLM is $\sim$3 times more powerful. Note that this is not a permanent feature of such models, and future models could have equally-sized components.

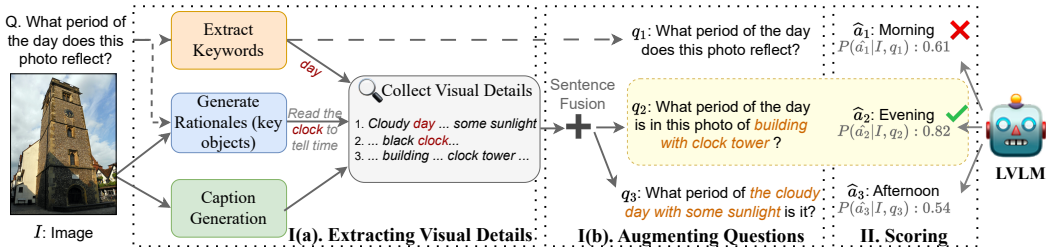

Figure 2: Schematic of REPARE for an image requiring implicit reasoning from A-OKVQA. We first extract keywords, captions, and rationales from the image conditioned on the question, which are used to identify important objects (e.g., day and clock). We query an LVLM about these objects to collect visual details in I(a), that are fused into the original question to produce, in this case, $n = 3$ candidates (I(b)). Lastly, we score and select from candidates using LVLM's answer confidence (II).

Changpinyo et al., 2022; Guo et al., 2023; Berrios et al., 2023). Given the higher performance of projection-based models on VQA, we focus our efforts on the former class.

LLMs can be used for multimodal chain-of-thought (CoT) reasoning (Zhang et al., 2023c); while we use forms of CoT in REPARE, the overall framework differs from CoT. Firstly, CoT is typically open-ended, whereas we follow a principled set of modules, which we validate individually in Sec. 4.1. Secondly, while CoT is generally useful on large models over 100 billion parameters (Magister et al., 2023; Wei et al., 2022a), REPARE can also be applied to LLMs like Flan-T5 (which do not generally benefit from CoT) without any modifications to the model (Wei et al., 2022a).

Underspecification and Ambiguity. Underspecification and ambiguity are well-studied within both NLP and linguistics (Schutze, 1995; Futeral et al., 2022; Berzak et al., 2015; Min et al., 2020; Rasmussen & Schuler, 2020). In the multimodal context, Pezzelle (2023) emphasizes underspecification as a significant source of errors in VL tasks – we develop REPARE as a concrete solution to address underspecification by adding visual information. Similarly, Bhattacharya et al. (2019) find underspecification to be a factor contributing to annotator disagreement in VQA, while Stengel-Eskin et al. (2023) focus on ambiguity in VQA and propose a rephrasing method for disambiguation. Unlike REPARE, their method relies on access to gold answers and involves further model training.

Prompt Editing. Both LLMs and LVLMs suffer from inherent randomness and sensitivity to choice of training examples, instructions, and prompt template (Zhao et al., 2021; Min et al., 2022; Lu et al., 2022; Awal et al., 2023) in zero-shot and few-shot settings. As a result, several works aim to search for better prompts via gradient-based (Shin et al., 2020; Gao et al., 2021; Jia et al., 2022; Khattak et al., 2023) or gradient-free methods (Sun et al., 2022; Deng et al., 2022; Prasad et al., 2023; Zhang et al., 2023b). However, existing gradient-based methods can be computationally expensive (Sung et al., 2022a), are infeasible for gated models accessible only via APIs, and are often uninterpretable (Khashabi et al., 2022). On the other hand, existing gradient-free methods are primarily designed for language-only models and select the best prompt based on scores (Liu et al., 2022; Prasad et al., 2023) or a learned policy (Deng et al., 2022; Zhang et al., 2023b) using a labeled training set. In contrast, REPARE directly addresses underspecification in VL tasks by making targeted edits to the question using gradient-free, instance-level edits without any train set.

3 METHODOLOGY

In this section, we describe the overall pipeline of our method: **Rep**hrase, **A**ugment and **Re**ason (REPARE). Broadly, REPARE consists of two stages: (I) *generating rephrased and augmented question candidates* and (II) *candidate selection*. The first stage yields n modified question candidates, incorporating visual information, and information from rationales using the underlying LVLM. We then use a selection module to identify the best candidate. Note that in all cases, selected questions should preserve the *intent* of the original question while making it easier for the model to answer. Fig. 2 provides a detailed illustration of our REPARE pipeline in action.

3.1 GENERATING REPHRASED AND AUGMENTED QUESTION CANDIDATES

Stage I(a): Extracting Visual Details from Captions and Rationales. To augment the question with pertinent visually-grounded details, we focus on extracting all relevant information from the image, conditioned on the question.

(i) *Salient Question Entities*: Intuitively, entities mentioned in the question provide vital information about the expected answer. To implement extract key entities from the question, we use an off-the-shelf keyword extraction system (Rose et al., 2010). For instance, it extracts *"day"* from the question in Fig. 2.

(ii) *Information from Rationales*: Answering complex questions can often require world knowledge and implicit reasoning skills (Schwenk et al., 2022). To incorporate this, we sample rationales from the LVLM, which we use to identify relevant objects and features in the image (Chowdhery et al., 2022; Zhang et al., 2023c). This allows REPARE to identify what features might be worth focusing on.[3] For instance, in Fig. 2, the model might extract the clock on the top of the building as an important feature in determining the time of day, based on a rationale like *"Clocks can tell time, so read the clock to determine the time of day."*

(iii) *General Information from Image Captions*: Questions may be underspecified to the extent that they do not contain any salient entities (e.g., *"where are we at"* in Fig. 1). Thus, we also prompt the LVLM to generate a detailed caption for the image. This allows us to capitalize on LVLMs' asymmetric abilities: they excel at image captioning (Alayrac et al., 2022; Tsimpoukelli et al., 2021; Zhu et al., 2023b), and can generate detailed captions (Zhu et al., 2023b; Xie et al., 2022). For example, in Fig. 2, the captioning model might generate a caption like *"A tall, stone building with a clock tower on top on a cloudy day"*.

After identifying salient objects and entities from (i) and (ii), we prompt the LVLM to obtain pertinent details about them based on the image. We add this list to the image captions from (iii) to get the input for for the next stage. We describe the implementation of this stage in detail in Appendix A.3.

Stage I(b): Rephrasing and Augmenting the Question. Drawing on work in sentence fusion (Geva et al., 2019; Lebanoff et al., 2020), we leverage the frozen LLM component of the LVLM to incorporate fine-grained details into the question. We combine all the extracted details into a single prompt, and generate $n-1$ modified question candidates, yielding a total of n candidates including the original question (see stage I(b) in Fig. 2). To prevent significant alteration of the question's meaning (especially for yes/no questions), we use an off-the-shelf natural language inference model (Laurer et al., 2022) to discard any candidates that contradict the original question. After generating n question candidates, we prompt the LVLM to answer each question, leading to n question-answer (QA) pairs. We use $n = 5$ as default in all our experiments and discuss the impact of increasing n as well as using the full LVLM for sentence fusion in Appendix A.5. Note that all our prompts used within REPARE or for VQA *do not* contain any annotated examples from any VQA dataset (zero-shot setting). Further details and all prompts can be found in Appendix A.2.

3.2 QUESTION SELECTION

To select the final QA pair from I(b), REPARE requires a way of scoring the n QA candidates generated using the modules above, in order to choose the QA pair most likely to improve accuracy.

Stage II: Confidence-based Selection. As discussed in Sec. 2, most prompt search methods require a labeled dataset to learn a scoring model or a selection policy. In our setting, we perform instance-level edits, meaning that such a supervised scoring scheme would require access to additional annotated data. Therefore, consistent with Liu et al. (2021), at inference time we compute an unsupervised score by utilizing the LLM's ability to self-assess the quality of its generations (Rae et al., 2021; Srivastava et al., 2023; Kadavath et al., 2022).[4] Following Kadavath et al. (2022), we use the LVLM's confidence in generating a proposed answer $\hat{a}_i$ conditioned on the image I and question candidate q_i to select candidate q' (and its corresponding answer $\hat{a}'$) for subsequent evaluation:

$$\text{score}(q_i, \hat{a}_i) = P_{\text{LVLM}}(\hat{a}_i | I, q_i); \qquad q', \hat{a}' = \underset{i \in [1,n]}{\text{argmax}}(\text{score}(q_i, \hat{a}_i))$$

Oracle Setting. As an upper-bound, we also explore an 'oracle' setting in which we have access to the (gold) annotated answer from the dataset. In this setting, we select the candidate that yields the

[3] Note that in the scope of this work, we do not address the veracity and utility of generated rationales which is relatively harder to judge using the same underlying model (Pruthi et al., 2022; Saha et al., 2023). Moreover, for one of the LVLMs (MiniGPT-4) using larger and more powerful LLMs, we prompt the model to generate an explanation in the *common* zero-shot VQA prompt for generating answers. Refer to Appendix A.2 for details.

[4] While past work has found LLMs to be overconfident on a variety of tasks (Mielke et al., 2022; Lin et al., 2022; Zhou et al., 2023; Stengel-Eskin & Van Durme, 2023), this does not impact our results, as we choose q' based on the LLM's *relative* confidence. For more details, we refer readers to Appendix A.5.

correct answer (in case of ties, we perform random selection). This gives us the maximum possible performance of REPARE for a fixed number of candidate questions n, discussed further in Sec. 4.

3.3 EXPERIMENTAL SETUP

Vision Language Models. We use three recent state-of-the-art LVLMs: BLIP-2 (Li et al., 2023), MiniGPT-4 (Zhu et al., 2023b), and LLaVA-1.5 (Liu et al., 2023a). At a high level, the model architecture comprises of an image encoder (Radford et al., 2021; Fang et al., 2023) and an LLM (Chung et al., 2022; Chiang et al., 2023) (both frozen) connected by a relatively small trained transformer model called the Q-former (Li et al., 2023). The Q-former acts as a bridge, facilitating information flow between the image encoder and the LLM, resembling an adapter (Houlsby et al., 2019; Sung et al., 2022b). Beginning with image-to-text pre-training, the Q-former extracts key visual details and then connects to the LLM using a fully-connected layer to project query embeddings into the embedding space of the LLM. Note that while BLIP-2 uses an encoder-decoder-based LLM (Flan-T5), MiniGPT-4 and LLaVA-1.5 use Vicuna with a decoder-only architecture (details in Appendix A.3).

VQA Datasets and Metrics. We use the VQAv2 dataset (Goyal et al., 2017) for general visual understanding. To specifically capture underspecification due to lack of reasoning or world-knowledge, we use the A-OKVQA dataset (Schwenk et al., 2022) containing image-question pairs that require broader commonsense and world knowledge to answer. A-OKVQA has two settings: (i) directly generating the answer (direct), and (ii) 4-way multiple choice (MC). Since the test sets of these benchmarks are not publicly available, we report performance on the validation sets (unless mentioned otherwise). Lastly, we also evaluate on the challenging VizWiz benchmark (Gurari et al., 2018) consisting of real-life information-seeking questions about (often low-quality) images sourced from visually-impaired people. While developing REPARE, we sample a small set of data points from the train set of the datasets to form our dev set. In the "direct answer" setting, we use the standard soft VQA evaluation metric for VQAv2, VizWiz, and A-OKVQA (Antol et al., 2015). In A-OKVQA's MC setting, we use accuracy. See Appendix A.1 for further dataset details.

4 RESULTS AND ANALYSIS

In this section, we present the results of our experiments. First, we establish the effectiveness of the REPARE framework in Sec. 4.1. Then, in Sec. 4.2, we quantitatively distinguish REPARE from simply paraphrasing the question. Furthermore, we provide quantitative analysis of outputs from REPARE, addressing semantic complexity (in Sec. 4.3). Lastly, in Sec. 4.4, we show that REPARE leverages the asymmetric strength of the LLM in an LVLM, allowing the LLM to perform more of the task without eliminating the need for the image.[5] Note that all improvements in this paper are reported as *absolute* percentage increase.

4.1 OVERALL EFFECTIVENESS OF REPARE

Main Results. Our main results are presented in Table 1. When compared to the original questions, using questions after applying REPARE increases the overall zero-shot accuracy of BLIP-2 by 3.85%, of MiniGPT-4 by up to 3.02%, and that of LLaVA-1.5 by 1.14% on VQAv2. On the A-OKVQA dataset, where answering the question may require a combination of world knowledge and reasoning skills, we show that REPARE improves the zero-shot performance of BLIP-2, MiniGPT-4, and LLaVA-1.5 models by up to 5.47%, 6.41%, and 3.63% respectively, when directly generating the answer. In the multiple-choice setting with the MiniGPT-$4_{\text{Vicuna 7B}}$ model, this improvement can be as high as 21.54%. Moreover, on the challenging VizWiz dataset, REPARE improves performance by 7.94%, 3.46%, and 2.39% points with MiniGPT-4, BLIP-2, and LLaVA-1.5 models. Furthermore, using gold answers in the oracle setting, we establish empirical upper bounds for REPARE in Table 1. On BLIP-2, REPARE can yield up to 9.84% across both datasets, while using MiniGPT-4, we can obtain a maximum oracle improvement of 14.41% and 33.94% on the A-OKVQA dataset in the direct and multiple-choice settings, respectively. Lastly, REPARE in oracle setting yields up to 7.61% accuracy improvements on the VizWiz dataset. This demonstrates REPARE's efficacy on VQA datasets with different LVLM architectures varying in size and underlying LLM.

Design Ablations. In Sec. 3, we described various design choices made to develop REPARE. In Table 2, we evaluate the effectiveness of different components within REPARE on our dev splits.

[5] We refer readers to Sec. 5 for a broader discussion of asymmetric strength and ability.

Method	VQAv2				A-OKVQA		VizWiz
	Overall	Y/N	Num.	Other	Direct	MC	Overall
MiniGPT-4$_{\text{Vicuna 7B}}$	51.47	71.03	27.13	42.75	27.51	41.66	29.87
+ REPARE	54.49$_{\pm 1.44}$	75.77$_{\pm 2.01}$	32.58$_{\pm 2.29}$	43.79$_{\pm 0.72}$	33.23$_{\pm 2.31}$	63.20$_{\pm 5.64}$	37.81$_{\pm 4.26}$
+ REPARE (Oracle)	59.66$_{\pm 2.93}$	86.56$_{\pm 6.12}$	38.86$_{\pm 4.57}$	44.32$_{\pm 1.05}$	41.92$_{\pm 4.86}$	75.60$_{\pm 7.36}$	50.47$_{\pm 5.32}$
MiniGPT-4$_{\text{Vicuna 13B}}$	61.98	82.76	39.83	51.71	41.53	56.41	44.18
+ REPARE	64.03$_{\pm 1.26}$	83.54$_{\pm 1.12}$	47.50$_{\pm 3.17}$	53.34$_{\pm 0.98}$	47.94$_{\pm 2.25}$	62.18$_{\pm 4.48}$	51.33$_{\pm 3.79}$
+ REPARE (Oracle)	68.35$_{\pm 3.62}$	89.33$_{\pm 4.28}$	51.41$_{\pm 5.67}$	56.30$_{\pm 2.27}$	54.20$_{\pm 5.18}$	80.67$_{\pm 8.26}$	64.27$_{\pm 4.68}$
BLIP-2$_{\text{Flan T5-xl}}$	62.58	83.99	38.63	52.91	41.86	73.89	59.63
+ REPARE	66.43$_{\pm 1.21}$	89.94$_{\pm 3.27}$	48.56$_{\pm 2.85}$	52.94$_{\pm 0.61}$	44.87$_{\pm 1.34}$	77.20$_{\pm 1.45}$	62.38$_{\pm 1.14}$
+ REPARE (Oracle)	72.42$_{\pm 3.49}$	94.26$_{\pm 4.73}$	50.67$_{\pm 5.06}$	61.25$_{\pm 2.94}$	49.20$_{\pm 2.38}$	81.14$_{\pm 3.61}$	65.20$_{\pm 2.65}$
BLIP-2$_{\text{Flan T5-xxl}}$	65.08	85.15	39.91	56.19	41.86	76.59	62.81
+ REPARE	68.92$_{\pm 1.36}$	90.54$_{\pm 3.18}$	43.30$_{\pm 2.41}$	58.96$_{\pm 0.87}$	47.33$_{\pm 1.56}$	**79.23**$_{\pm 1.29}$	**66.27**$_{\pm 1.87}$
+ REPARE (Oracle)	74.05$_{\pm 3.26}$	94.56$_{\pm 4.37}$	54.40$_{\pm 4.72}$	63.36$_{\pm 2.84}$	55.67$_{\pm 2.19}$	82.80$_{\pm 2.36}$	70.13$_{\pm 2.49}$
LLaVA-1.5$_{\text{Vicuna-7B}}$	76.21	91.83	58.27	68.85	62.56	77.38	57.07
+ REPARE	**77.35**$_{\pm 0.42}$	**92.64**$_{\pm 0.34}$	**59.92**$_{\pm 0.51}$	**70.12**$_{\pm 0.76}$	**66.19**$_{\pm 1.53}$	78.21$_{\pm 0.37}$	59.46$_{\pm 0.92}$
+ REPARE (Oracle)	79.84$_{\pm 1.03}$	94.39$_{\pm 0.94}$	62.07$_{\pm 1.25}$	73.28$_{\pm 1.31}$	70.17$_{\pm 2.49}$	80.75$_{\pm 1.27}$	62.48$_{\pm 1.61}$

Table 1: Comparison of baseline zero-shot accuracy (%) and REPARE on VQAv2, A-OKVQA and VizWiz. We run REPARE for $n = 5$ and average performance across 3 random seeds to account for randomness in generating question candidates in Sec. 3.1. We highlight the oracle performance with REPARE using gold answers. The overall best numbers for each dataset are in bold, and the highest numbers for each model are underlined.

- *Importance of Rationales, Captions, and Question Entities*: We measure the utility of details about objects mentioned in the: (i) original question, (ii) image caption, and (iii) rationales, by re-running REPARE with BLIP-2 using *all but one* type of object descriptions. From Table 2, we observe that excluding rationales, captions, or question entities adversely impacts zero-shot performance, with the largest drop in accuracy occurring when rationales are not utilized.
- *Impact of Removing Visual Tokens during Fusion*: Next, we explore the impact of including visual tokens, projected onto the LM, in augmenting the question with visual details in Stage I(b). This involves performing the same sentence fusion task using the entire LVLM, while retaining the image embedding in the input to the frozen LM (refer to Table 2). Our findings reveal that the image embedding can serve as a distraction to the language model when rephrasing the question, resulting in up to a 3.1 point drop in overall accuracy (see qualitative examples in Appendix A.5).
- *Design of Scoring Function*: Lastly, we examine our scoring function described in Sec. 3.2. To ablate the scoring method, we run REPARE but with candidates based on the likelihood of the question *alone*, i.e. $\text{score}(q_i) = P_{\text{LVLM}}(q_i|I)$ instead of $\text{score}(q_i, \hat{a}_i) = P_{\text{LVLM}}(\hat{a}_i|I, q_i)$. Table 2 shows that using question likelihood instead of the answer confidence yields a small drop in the downstream performance by at least 1.09% (further ablations in Appendix A.5).

4.2 REPARE ADDS SEMANTIC INFORMATION TO ADDRESS UNDERSPECIFICATION

Comparison with Paraphrasing in Oracle Setting. Following past work on leveraging paraphrases to improve QA (Dong et al., 2017), we experiment with a paraphrastic baseline, where we simply paraphrase the question using Pegasus, a strong off-the-shelf model (Zhang et al., 2020).

Table 3 shows that paraphrasing the question leads to major improvements over the zero-shot setting under oracle selection (described in Sec. 3.2). For VQAv2, BLIP-2's performance increases from 62.58% to 70.99% and for A-OKVQA it improves from 73.89% to 79.91% in the multiple choice setting. This indicates that BLIP-2 and its underlying LLM, Flan-T5 are brittle to the phrasing of the question, i.e., without altering the information or meaning of the question, a paraphrased question candidate may yield a higher VQA score (Webson & Pavlick, 2021).

Comparison with Paraphrasing during Inference. If the modifications from REPARE were purely cosmetic rewrites, then REPARE and a paraphrastic baseline should have roughly the same performance during inference. Table 3 demonstrates that selecting from paraphrased question candidates without access to gold answers (oracle) presents a challenge. In fact, in 2 out of 3 settings, opting for paraphrased questions results in *lower* performance compared to using the original ques-

Method	VQAv2	A-OKVQA
REPARE	**67.28**	**45.01**
w/o Rationales	64.57	43.15
w/o Caption	66.04	44.36
w/o Question Entity	65.89	44.62
w/ I Embeddings in Fusion	63.18	42.38
w/ score $= P_{\text{LVLM}}(q_i\|I)$	65.47	43.92

Table 2: Ablation of design choices in REPARE using BLIP-2 on our dev splits (direct answers).

Method	VQAv2	A-OKVQA	
	Overall	Direct	MC
Baseline (BLIP-2)	62.58	41.86	73.89
Paraphrase Oracle	70.99	46.66	79.91
REPARE Oracle	**72.42**	**49.24**	**81.14**
Paraphrase Selection	62.91	40.23	73.57
REPARE Selection	**66.43**	**44.87**	**77.20**

Table 3: Comparison of REPARE (using BLIP-2) with paraphrasing questions in the oracle setting and unsupervised candidate selection.

tions by up to 1.63%. Therefore, although some paraphrased questions may elicit correct answers, choosing them solely based on the model's confidence yields poor results. In contrast, questions generated by REPARE show a distinct pattern: not only do these questions outperform paraphrased questions in the oracle setting, but they are also more easily chosen by the unsupervised scoring function. This indicates that incorporating additional semantic information from both images and rationales in REPARE simultaneously makes questions *easier to answer* as well as *easier to select*.

4.3 ANALYSIS OF INCREASED COMPLEXITY IN REPARE'S QUESTIONS

In Sec. 1, we highlight underspecification as a source of errors in VL tasks like VQA. In Table 1, we empirically show that REPARE enhances VQA accuracy across datasets and LVLM architectures. Here, we analyze the questions generated by REPARE and compare them against original questions to confirm that the rephrased questions are in fact more complex, i.e., *less underspecified*. We present quantiative results from two complexity metrics; see Appendix A.4 for qualitative examples.

Complexity Metrics. Qualitatively, we find that REPARE questions have increased syntactic and semantic complexity (cf. Table 6). We quantify this with two common complexity metrics: average dependency distance (ADD) and idea density (ID), implemented using BlaBla toolkit (Shivkumar et al., 2020) and Stanza (Qi et al., 2020). *Average Dependency Distance* (ADD) measures the *syntactic complexity* of sentences by calculating the average linear distance between each token and its parent node in a syntactic parse. It is commonly used to measure syntactic complexity (Gibson et al., 2000; Oya, 2011; Liu et al., 2017). ADD ranges on $[0, \inf)$ with a higher score indicating more complexity. *Idea Density* (ID) is the sum of the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words (Boschi et al., 2017). It is commonly used as a measure of *semantic complexity* (Chand et al., 2012; Kemper, 1992). ID ranges between $[0, 1]$ and higher scores indicate more complexity.

Dataset	Type	ADD	ID
A-OKVQA	Original	25.40	0.282
	REPARE	32.81	0.299
VQAv2	Original	17.87	0.258
	REPARE	29.52	0.296

Table 4: Complexity measures for questions before and after REPARE.

Quantitative Complexity Analysis. Our quantitative results can be seen in Table 4, where we compute ADD and ID for a subset of 100 instances from the official validation set. Here, we use BLIP-2 as the backbone for REPARE. Compared to the original questions, both complexity measures are higher for REPARE across models and datasets. This indicates that REPARE adds syntactic complexity and semantic content to the questions; which in turn suggests that the rephrased questions are less underspecified. For example, a REPARE question like *"Why would you use this suitcase packed on both sides?"* from Table 6 has more modifiers than the original, *"Why would you use this bag?"*, leading to a higher ID score. It also has a more complicated syntactic structure, with nested modifiers (*"suitcase packed on both sides"*) leading to a higher ADD.

4.4 REPARE LEVERAGES VL INTERACTION TO IMPROVE PERFORMANCE

We further explore the *asymmetric strength* hypothesis (discussed in Sec. 1),[5] which could explain the improvements seen in Table 1. Specifically, we examine how REPARE's addition of visual information to the question allows the LVLM's LLM component to do more of the heavy lifting in the QA task. We test the performance of the original and REPARE questions *without* the image in the input, i.e., to what extent the constituent LLM alone can answer each question cor-

rectly. If REPARE leverages the strength of the LLM well, we should expect the LVLM's LLM-only performance to increase when using REPARE. In Table 5, we evaluate this hypothesis using BLIP-2 as the underlying model. First, we observe that the *image is crucial* to good performance; in all settings, BLIP-2's LLM-only accuracy is quite low. Furthermore, REPARE questions improve in the LLM-only setting, indicating that modified questions take better advantage of the LLM's QA strength (cf. rows 3 and 4). Note the substantial gap of $\sim$25% between settings with and without the image for REPARE (cf. rows 2 and 4), which indicates that the rephrased question is *complementary* to the image, i.e., that REPARE does not make questions trivial to answer with just an LLM. Finally, when using just the LLM as the QA model, we find that adding the caption or extracted image details from stage I(a) of REPARE along with the original question improves performance over the original question alone; however, these details do not make up for the lack of the image (c.f. rows 2, 5, and 6 in Table 5). Thus, REPARE improves LVLM performance via both vision-language interaction *and* leveraging the LLM.

	LVLM Setting	VQAv2	A-OKVQA	
		Overall	Direct	MC
[1]	Img + Orig. Q	60.29	41.72	72.56
[2]	Img+ REPARE Q	67.28	45.01	78.43
[3]	Orig. Q (LLM-only)	32.84	15.93	45.20
[4]	REPARE Q (LLM-only)	40.53	20.33	54.21
[5]	Caption + Orig. Q	52.88	27.64	65.80
[6]	REPARE I(a) + Orig. Q	54.31	30.27	66.60

Table 5: BLIP-2's LLM-only vs. full model performance on original and rephrased questions.

5 DISCUSSION AND CONCLUSION

Asymmetric Strength and Ability. As alluded to in Sec. 1, REPARE is based on two assumptions about existing LVLMs. First, existing LVLMs have larger LLMs than vision components, i.e. they have *asymmetric strength*. Thus, moving some of the burden of the VQA task onto the LLM improves performance, as shown in Sec. 4.4. However, we also assume that the image is still helpful in answering the question – this is also borne out in Sec. 4.4, where visual information still improves the model. This differentiates our work from recent work like Berrios et al. (2023) and Hu et al. (2022) which translate an image into text descriptions and then apply a language-only model for VL tasks, i.e. also make use of asymmetric strength, but not of the image. Our work also holds greater promise for capturing fine-grained visual details, which can be challenging to describe linguistically. Second, REPARE also relies on the asymmetric zero-shot abilities of individual LVLMs. While there is a large gap in QA between zero-shot LVLMs and fine-tuned, task-specific models, LVLMs are competitive at image captioning. We can use this to our advantage, harnessing captions to improve the question. Similarly, while LVLMs may not be able to implicitly reason about the image during QA, their LLMs can extract useful rationales and fuse them into the question.

Redundancy and Language Bias. Qualitatively, much of the information in Table 6 may appear redundant to humans who can perceive the entire image and hone in details already given in the image. It is worth noting that past work has found that humans tend to over-specify when describing visual scenes (Ford & Olson, 1975; Sonnenschein, 1985; Pechmann, 1989; Koolen et al., 2011). In other words, redundancy in descriptions or questions is not uncommon, and may in fact benefit the model. VQA datasets can suffer from language bias, where many questions can be answered correctly without access to the image (Goyal et al., 2017). The analysis in Sec. 4.4 indicates that REPARE questions have stronger LLM-only (i.e., language-only) performance. However, note that the information that gives REPARE questions their higher performance is *extracted from the image* using the same underlying LVLM. Thus, a comparison to a language-only bias here is not entirely accurate, since rephrased questions contain information sourced from the image.

Limitations. One limitation of our method is cost: rather than answering a question directly, we generate several question candidates and then select one. We note, however, that other multi-step approaches, including Chain-of-Thought (Wei et al., 2022b), or exploration-guided reinforcement learning, and search methods also increase the number of tokens and inference steps, and our cost scales linearly in the number of candidates. Note also that while strategies like CoT can help with rationale-style reasoning in particular, they are harder to apply in most existing LVLMs (partly due to the size of LLM components in existing LVLMs, which is typically significantly less than 100B parameters). Addressing underspecification alone is not a cure-all for solving VQA or broader visual understanding tasks. Underlying dataset issues, such as low-quality images and inaccurate human annotations (Bhattacharya et al., 2019), can still prevent models from achieving high accuracy.

ETHICS STATEMENT

Instructions are a useful tool for conveying extrinsic information to LLMs. However, they can also be misused intentionally or unintentionally (Weidinger et al., 2021) in order to alter model outputs to elicit harmful, biased and problematic content. Being based on LLMs, LVLMs are succeptible to similar misuse via targeted instructions or questions. The intended use of REPARE is to obtain modifed questions that work well for LVLMs and help improve model performance for the given instance without significanlty altering the intended meaning; this is orthogonal to whether the original question displays a malicious intent, which is a more general issue that applies to all LLMs/LVLMs. Additionally, in our work we use images and questions from VQA and A-OKVQA; these datasets have been vetted for quality in the past Lin et al. (2014); Goyal et al. (2017); Schwenk et al. (2022) but inappropriate or offensive queries and images could remain since they are quite large. To mitigate the risk of offensive, malicious, or inappropriate questions being generated by REPARE, we manually examined a subsample of 250 generated outputs from REPARE and verified that the generated questions do not display a malicious or offensive intent.

ACKNOWLEDGEMENTS

We thank Jaemin Cho, Peter Hase, Nithin Sivakumaran, David Wan, Jaehong Yoon, and Shoubin Yu for their valuable feedback and inputs for the paper. This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, DARPA Machine Commonsense (MCS) Grant N66001-19-2-4031, ARO Award W911NF2110220, and ONR Grant N00014-23-1-2356. The views contained in this article are those of the authors and not of the funding agency.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023.

William Berrios, Gautam Mittal, Tristan Thrush, Douwe Kiela, and Amanpreet Singh. Towards language models that can see: Computer vision through the lens of natural language. *arXiv preprint arXiv:2306.16410*, 2023.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. Do you see what I mean? Visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1477–1487, 2015.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4271–4280, 2019.

Veronica Boschi, Eleonora Catricala, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F Cappa. Connected speech in neurodegenerative language disorders: a review. *Frontiers in psychology*, 8:269, 2017.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pp. 287–318. PMLR, 2023.

Vineeta Chand, Kathleen Baynes, Lisa M Bonnici, and Sarah Tomaszewski Farias. A rubric for extracting idea density from oral language samples. *Current protocols in neuroscience*, 58(1): 10–5, 2012.

Soravit Changpinyo, Doron Kukliansy, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1947–1963, 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3369–3391, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.222. URL https://aclanthology.org/2022.emnlp-main.222.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 875–886, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1091. URL https://aclanthology.org/D17-1091.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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

William Ford and David Olson. The elaboration of the noun phrase in children's description of objects. *Journal of Experimental Child Psychology*, 19(3):371–382, 1975.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. *arXiv preprint arXiv:2212.10140*, 2022.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.295. URL https://aclanthology.org/2021.acl-long.295.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. 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)*, pp. 3443–3455, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1348. URL https://aclanthology.org/N19-1348.

Edward Gibson et al. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000:95–126, 2000.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10877, 2023.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019. URL http://proceedings.mlr.press/v97/houlsby19a.html.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. Improved lexically constrained decoding for translation and monolingual rewriting. 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)*, pp. 839–850, 2019.

Yushi* Hu, Hang* Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Susan Kemper. Language and aging. *The handbook of aging and cognition*, 1992.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3631–3643, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.266. URL https://aclanthology.org/2022.naacl-main.266.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. Effects of scene variation on referential overspecification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.

Moritz Laurer, W v Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying–addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli, 2022. URL https://osf.io/wqc86/.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. Understanding points of correspondence between sentences for abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 191–198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-srw.26. URL https://aclanthology.org/2020.acl-srw.26.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Haitao Liu, Chunshan Xu, and Junying Liang. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of life reviews*, 21:171–193, 2017.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021. URL https://arxiv.org/abs/2101.06804.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4300–4312, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.340. URL https://aclanthology.org/2021.naacl-main.340.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. Reframing instructional prompts to gptk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022. URL https://arxiv.org/abs/2109.07830.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Masanori Oya. Syntactic dependency distance as sentence complexity measure. In *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics*, volume 1, 2011.

Thomas Pechmann. Incremental speech production and referential overspecification. *Linguistics*, 1989.

Sandro Pezzelle. Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12098–12112, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.675. URL https://aclanthology.org/2023.acl-long.675.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1314–1324, 2018.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. Grips: Gradient-free, edit-based instruction search for prompting large language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3827–3846, 2023.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, Liwei Jiang, Yejin Choi, and Chandra Bhagavatula. ClarifyDelphi: Reinforced clarification questions with defeasibility rewards for social and moral situations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11253–11271, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.630. URL https://aclanthology.org/2023.acl-long.630.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 101–108, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Nathan Ellis Rasmussen and William Schuler. A corpus of encyclopedia articles with logical forms. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pp. 1–20, 2010.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*, 2023.

Hinrich Schutze. *Ambiguity in language learning: Computational and cognitive models*. Stanford University, 1995.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.346. URL https://aclanthology.org/2020.emnlp-main.346.

Abhishek Shivkumar, Jack Weston, Raphael Lenain, and Emil Fristed. Blabla: Linguistic feature extraction for clinical analysis in multiple languages. *Proc. Interspeech 2020*, pp. 2542–2546, 2020.

Susan Sonnenschein. The development of referential communication skills: Some situations in which speakers give redundant messages. *Journal of Psycholinguistic Research*, 14(5):489–508, 1985.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Elias Stengel-Eskin and Benjamin Van Durme. Calibrated Interpretation: Confidence Estimation in Semantic parsing. *Transactions of the Association for Computational Linguistics*, 2023. doi: https://arxiv.org/pdf/2211.07443.pdf.

Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. Why did the chicken cross the road? rephrasing and analyzing ambiguous questions in VQA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10220–10237, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.569. URL https://aclanthology.org/2023.acl-long.569.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. *arXiv preprint arXiv:2201.03514*, 2022.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005, 2022a.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2022b.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WtmMyno9Tq2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*, 2021. URL https://arxiv.org/abs/2109.01247.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022b.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021. URL https://arxiv.org/abs/2112.04359.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. Visual clues: Bridging vision and language foundations for image paragraph captioning. *Advances in Neural Information Processing Systems*, 35:17287–17300, 2022.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3081–3089, 2022.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pp. 11328–11339. PMLR, 2020. URL http://proceedings.mlr.press/v119/zhang20ae/zhang20ae.pdf.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023a.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E. Gonzalez. TEMPERA: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=gSHyqBijPFO.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023c.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pp. 12697–12706. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhao21c/zhao21c.pdf.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *arXiv preprint arXiv:2302.13439*, 2023.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023a.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023b.

A APPENDIX

A.1 DATA

We use three datasets, VQAv2 (Goyal et al., 2017), A-OKVQA (Schwenk et al., 2022), and VizWiz (Gurari et al., 2018). VQAv2 is an extension of the original VQA dataset (Antol et al., 2015), which incorporates similar images yielding different answers to the same question. This augmentation doubles the number of image-question pairs, emphasizing the reliance on visual information for accurate answers. While VQA questions are open-ended, the answer vocabulary is relatively limited in size (10M), consisting of mostly one-word responses. In VQAv2, each example is associated with 10 ground-truth answer labels provided by different human annotators. On the other hand, the A-OKVQA dataset is smaller (25K questions in total) but is more challenging. Similar to VQAv2, in the direct answer setting, 10 human annotated 1-2 word answers are provided for each question. The multi-choice setting comes with 4 options along with the index of the correct option. Lastly, the VizWiz dataset contains 32.8K information-seeking questions asked by visually-impaired people based on images clicked on mobile devices. This dataset can be challenging as the images are often blurred, under/over-exposed, or rotated. During the design and analysis of REPARE, we use a separate development set consisting of 5K, 1K, 500 randomly sampled image-question pairs from the train sets of VQAv2, A-OKVQA, and VizWiz respectively. For testing, we use the entire validation set; this corresponds to 214K examples for VQA, 1.1K examples for A-OKVQA, and 4.3K examples for VizWiz. We use the standard VQA metric for open-ended evaluation. According to this metric, a model-generated answer is deemed 100% accurate if at least 3 of the 10 annotators provided that exact answer.

$$\text{Accuracy}_{\text{VQA}} = \min\left(\frac{\text{\# humans that said ans}}{3}, 1\right)$$

The predicted answer is also pre-processed by lowercasing, converting numbers to digits, and removing punctuation/articles. Since LLMs generate free-form text, we constrain the answers using length-penalty of -1 during generation which encourages shorter answers that align better with human annotations.

A.2 PROMPTS

Table 11 contains an exhaustive list of all prompts used in REPARE for various models. Limited prompt engineering (2-3) trials were done for each prompt on our dev split. For sentence fusion, we use two hypothetical examples (note that these do not include images):

1. **Question**: What is the man wearing?; **Object**: man; **Detail**: he is standing on the sidewalk; **Modified Question**:What is the man who is standing on the sidewalk wearing?
2. **Question**: Are there any flowers?; **Object**: flowers; **Detail**: There is flowers are in a vase. The vase is blue in color and sitting on a table; **Modified Question**: Are there any flowers in the vase on the table?

A.3 EXPERIMENTAL DETAILS

Model Checkpoints. In Sec. 3.3, we use BLIP-2 with ViT-g frozen image encoder (1B parameters), and Flan-T5 XL with 3B model parameters. The pretrained Q-former is an encoder-only transformer model (Vaswani et al., 2017) that shares a similar architecture with BERT (Devlin et al., 2019) comprising of 107M parameters. MiniGPT-4 and LLaVA-1.5 are based on the BLIP-2 architecture with an addition VL pretraining. One key difference is that it uses the Vicuna family of LLMs. We experiment with the two official checkpoints with 7B and 13B model parameters. In Sec. 4.2, we use a popular Pegasus-based paraphrasing model available on HuggingFace (Wolf et al., 2020).[6] We also use an off-the shelf NLI model that achieves competent performance on a suite of NLI benchmarks Laurer et al. (2022).[7] In Sec. 3, we also use the `rake_nltk` python package.

[6]Link to Checkpoint: https://huggingface.co/tuner007/pegasus_paraphrase

[7]Link: https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

Figure 3: Example images and original questions for Table 6. Some questions (e.g., *"What is behind the boy"* are underspecified, while others refer to small objects in the image (e.g., *"What color wetsuit is he wearing?"*).

Stage I: Extracting Visual Details and Generating Candidates. As described in Sec. 3.1, we extract salient visual details from the image using 3 components described below in further detail.

(i) *Salient Question Entities:* Given only the question from a data instance, we use the keyword extraction tool from `rake_nltk` package to identify salient keywords mentioned in the question. For instance, in the example shown in Fig. 2, "day" is identified as the salient entity.

(ii) *Information from Rationales:* For this module, we use both the image and the original question for the data point and adopt a two step approach. First, we ask the model to generate an explanation for its answer. Next, based on the explanation and question, we ask the model to identify salient entities mentioned or used in the rationales. Refer to `Rationale (ii)` prompts listed in Table 11.

(iii) *Information from Captions:* We adopt the straightforward approach of using the `caption` prompts from Table 11 to generate the image caption using the LVLM.

We extract visual information about entities identified in steps (i) and (ii) using the LVLM by querying it using the `extraction of details` prompt listed in Table 11 for each entity separately. This information is concatenated in a prompt in the form of a bulleted list of format `[entity] : [details]`. To this prompt, we add the generic details from the image caption via an additional line: `image: [caption]`. This list of details along with the original question are added to the sentence fusion prompt from Table 11 along with two in-context examples mentioned in Appendix A.2 above which generates the modified question candidates by sampling multiple outputs (same LVLM call).

Text Generation in REPARE. To ensure the paraphrasing model generates a valid question ending with '?' we employ constrained decoding by setting a positive constraint on generating the '?' token (Post & Vilar, 2018; Hu et al., 2019). To ensure diverse samples in the sentence fusion stage (determines the diversity of question candidates) we use top-p sampling Holtzman et al. (2019) with $p = 0.95$. To sample rationales, we employ beam search with 5 beams and a temperature of 0.7. After generating question candidates, we filter out sentences that are not valid (do not end with a question mark) or are a verbatim repetition of the original question. Additionally, we also filter out contradictory generations as described in Sec. 3. We sample enough candidates such that we are left with n distinct candidates in the end. If this is not feasible, we repeat the original question in the candidate set to make up for the difference.

Candidate Selection via REPARE's Score (Stage II). We employ the `VQA` prompts mentioned in Table 11 to obtain answers for each question candidate. Typically, the answers (direct answers or option label) correspond to one word or token. In case where we are scoring multiple tokens (as in Sec. 3.2 or Sec. 4.1), we compute the length (number of tokens) normalized log-probabilities that are subsequently exponentiated to obtain probabilites. Note that we are only interested in the relative order, therefore, we can alternatively use log-probabilities to score and select candidates too.

A.4 Qualitative Examples and Analysis

RepARe questions exhibit an increased degree of specificity, with additional modifiers and fewer ambiguous references, e.g., *"the person riding the wave on the surfboard"* as opposed to *"he"* in the original. Even when questions are unambiguous, RepARe questions include reasoning and location clues. For example, a rephrased question like *"What time is on the clock at the top building?"* indicates which region of the image is important.

	Original	RepARe
VQA	Does the water have ripples?	Does the water have the small ripples around the boats?
	What time does the clock say?	What time is on the clock at the top of building?
	What is behind the boy?	What is behind the boy doing a trick on a skateboard?
	What color wetsuit is he wearing?	What color is the wetsuit of the person riding the wave on the surfboard?
A-OKVQA	Why would you use this bag?	Why would you use this suitcase packed on both sides?
	Which piece of clothing is unique to one person here?	What piece of clothing is unique to the man on the right?
	What will be built here one day?	What will be built at this construction site?
	The men facing the camera are all wearing what?	The men facing the camera are all wearing the same pair of sunglasses?

Table 6: Qualitative examples of original and RepARe generated questions for both datasets with BLIP-2 as the underlying model. For corresponding images, refer to Fig. 3.

Fig. 3 shows the images corresponding to the examples given in Table 6. Each image is paired with its original question as well as the rephrased question from RepARe.

Answers in Questions. In some cases, e.g. in the final A-OKVQA question about sunglasses in Table 6, the correct answer (*"sunglasses"*) is added to the question by RepARe. We first note that this is not an unfair advantage, since RepARe operates on the same information as the QA model (image and question), using the same LVLM. Any additional information in a RepARe question is extracted from captions and rationales, which are obtained in a realistic zero-shot test-time setting without any access to the gold answer. Similarly, RepARe's selection module does not use the gold answer in selection. Nevertheless, we report the percentage of times the correct answer is found in the RepARe question and not in the original question. Here, we use the A-OKVQA open-ended (direct) setting and the BLIP-2 model. In a random sample of 100 examples , we find that 7% of rewritten questions from RepARe contain a gold answer. This indicates that part of RepARe's advantage likely comes from an ability to extract the correct answer from the caption and rationale information, incorporate into a question candidate, and then select that candidate.

A.5 Additional Ablations

Impact of LVLMs on Generating Question Candidates. As mentioned in Sec. 3.1, we only used the underlying LLM to fuse or incorporate the extracted visual details into the given question. In Table 2 of Sec. 4.1, we quantitatively show that including the visual tokens, i.e., using the entire LVLM negatively impacts the overall RepARe pipeline and decreases downstream performance. To provide additional insights, Table 7 contains qualitative examples of fusion using only the LLM and the entire LVLM using the BLIP model. We observe that the image embeddings serve as a distraction to the LLM when performing a primarily linguistic task and the resultant question is often ill-formed and heavily dominated by the image caption and/or visual details.

Alternate ways of computing Answer Confidence. Kadavath et al. (2022) demonstrate that the self-evaluation ability of model is better in multiple-choice settings than settings in which the LM is required to directly generate the answer. Note that a multiple choice setting is also better specified, since the model is conveyed a set of options to choose from. For instance, if multiple plausible answers exist, only one would be mentioned in the options, indirectly communicating the type of intended response. This is reflected by the contrast in A-OKVQA accuracy (cf. Table 1) in direct and MC settings. In the direct answer setting, we compare computing model's answer confidence in

	Original	REPARE w/ LLM Fusion	REPARE w/ LVLM Fusion
VQA	Does the water have ripples?	Does the water have the small ripples around the boats?	Yes, there are ripples in the body of water where a group of boats are docked?
	What time does the clock say?	What time is on the clock at the top of building?	What time does the clock tower say in the building?
	What is behind the boy?	What is behind the boy doing a trick on a skateboard?	A person is doing a trick on a skateboard behind the boy in the middle of a parking lot in the city at night in front of a large building with a skyscraper in the background?
	What color wetsuit is he wearing?	What color is the wetsuit of the person riding the wave on the surfboard?	What color wetsuit is he riding in?
A-OKVQA	Why would you use this bag?	Why would you use this suitcase packed on both sides?	Why would you use this bag with a lot of items in it?
	Which piece of clothing is unique to one person here?	What piece of clothing is unique to the man on the right?	What piece of clothing is unique to one person: a man and a woman posing for a picture?
	What will be built here one day?	What will be built at this construction site?	A truck is driving down a street with construction cones and a construction site in the background, what will be built there one day?
	The men facing the camera are all wearing what?	The men facing the camera are all wearing the same pair of sunglasses?	The men facing the camera are all wearing the same thing, what is it?

Table 7: Qualitative comparison of generated question candidates with and without visual tokens in Stage II (sentence fusion) of REPARE. Corresponding images shown in Fig. 3.

two additional ways that Kadavath et al. show to be better calibrated. First, we compute True/False answer confidence by adding the following suffix to the VQA prompt:

```
Proposed Answer: [â_i].
Is the proposed answer true or false? (A) True, (B) False.
The proposed answer is:
```

Then, we use $P_{\text{LVLM}}(\text{True})$ as a substitute for answer confidence $P_{\text{LVLM}}(\hat{a}_i|I, q_i)$. Additionally, we also employ the strategy of showing the model multiple generated answers to estimate answer confidence. For this we take advantage of n different question candidates that yield different answers $\{\hat{a}_i\}_{i\in[1,n]}$. Hence, we add the following prefix to the VQA prompt:

```
Plausible Answers: [{â_i}_{i∈[1,n]}].
Proposed Answer: â_i.
Is the proposed answer true or false? (A) True, (B) False.
The proposed answer is:
```

We denote this setting as $P_{\text{LVLM}}(\text{True}|\{\hat{a}_i\}_{i\in[1,n]})$ and substitute this probability instead in the score function. The results are shown in Table 8. We find that all the implementations of LVLM's answer confidence yield comparable performance across datasets (<1 point difference). Since we use the same LVLM to rank various question candidates for a given image, the relative ordering of scores (Sec. 3.2) should not be significantly affected by the model's calibration. Post-hoc calibration methods such as Platt scaling (Platt et al., 1999) or isotonic regression (Zadrozny & Elkan, 2002) preserve relative ordering, so they would have no effect on the selection criterion.

REPARE score	VQA	A-OKVQA
$P_{\text{LVLM}}(\hat{a}_i\|I, q_i)$	67.28	45.01
$P_{\text{LVLM}}(\text{True})$	68.01	44.78
$P_{\text{LVLM}}(\text{True}\|\{\hat{a}_i\}_{i\in[1,n]})$	67.56	44.87

Table 8: Comparison of performance of REPARE with BLIP-2 using different score functions for computing answer confidence.

Imapct of increasing the number of candidates n in REPARE. In Fig. 4, we explore the impact of increasing the number of candidates in REPARE, i.e., n on its effectiveness at enhancing BLIP-2's VQA (direct) performance. We find that initially increasing from $n = 2$ to $n = 5$ leads to performance gains in both the inference and oracle settings for VQAv2 and A-OKVQA datasets. However, the gains saturate after $n = 10, 15$ for both datasets. In fact, during inference, wherein we select 1 out of n question candidates, we find the VQA accuracy gradually decreases at $n = 15$. This is expected, since increasing n allows for diverse candidates; however, selection from a very large pool of candidates (like $n = 15$) is more challenging, and REPARE's selection module is more likely to make a suboptimal choice – hence the growing gap between oracle and REPARE performance.

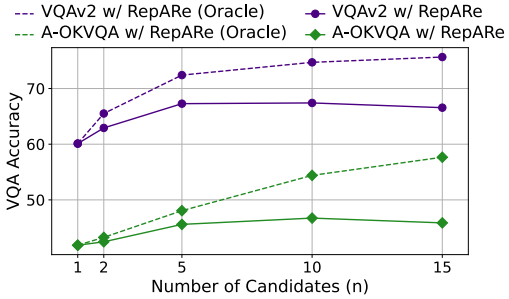

Figure 4: Trends in VQA performance of REPARE for different values of n.

Analysis with MiniGPT-4. In Sec. 4, we use BLIP-2 for conducting our analysis. However, BLIP-2 uses an encoder-decoder LLM while the remaining LVLMs (MiniGPT-4 and LLaVA-1.5) use Vicuna, which is a decoder-only LLM. In this section, we show that the choice of underlying LLM architecture does not impact the relative trends. In Tables 9 and 10, we repeat the analysis in Sec. 4.1 with MiniGPT-4 models corresponding to Tables 2 and 3 respectively. Our ablation study in Table 9 once again highlights the importance of each component of RepARe in improving VQA performance. Similarly, Table 10 reveals that even with MiniGPT-4 as the underlying LVLM, candidates generated by REPARE significantly outperform paraphrased question candiates when selected based on the model's answer confidence.

Method	VQAv2	A-OKVQA
REPARE	**57.74**	**31.23**
w/o Rationales	53.29	28.62
w/o Caption	56.49	29.51
w/o Question Entity	54.91	29.28
w/ I Embeddings in Fusion	54.49	28.97
w/ score $= P_{\text{LVLM}}(q_i\|I)$	56.46	29.19

Table 9: Ablation of design choices in REPARE using MiniGPT-$4_{\text{Vicuna 7B}}$ on our dev splits (direct answers).

Method	VQAv2	A-OKVQA	
	Overall	Direct	MC
Baseline (MiniGPT-4)	51.47	27.51	41.66
Paraphrase Oracle	57.41	39.83	73.68
REPARE Oracle	**59.66**	**41.92**	**75.60**
Paraphrase Selection	51.39	26.28	40.52
REPARE Selection	**54.49**	**33.23**	**63.20**

Table 10: Comparison of REPARE (MiniGPT-$4_{\text{Vicuna 7B}}$) with paraphrasing questions in the oracle setting and unsupervised candidate selection.

	Dataset	Setting	Prompt
BLIP-2	VQAv2	VQA Prompt	`Question: [Question] Short Answer:`

	A-OKVQA (MC)	VQA Prompt	`Question: [Question]`
			`Options: A. [Choice 1], B. [Choice 2], C. [Choice 3] , D. [Choice 4]`
			`Answer: Option`

	All	Caption	(Default, empty string)
		Rationale (i)	`[VQA Prompt] Explanation:`

		Rationale (ii)	`[LVLM Response for Rationale (i)]`
			`Question: [Question]`
			`Which all entities or objects from this image would I need to observe to answer this question?`

		Extraction of Details	`Question: What can you tell me about [entity] in this image?`
MiniGPT-4	VQAv2	VQA Prompt	`### Human: <Img> <ImageHere> </Img>### Human: Based on the image, answer the question below in preferably only 1 word.` `Question: [Question]`
	A-OKVQA	VQA Prompt (i)	`### Human: <Img> <ImageHere> </Img>### Human: Based on the image, answer the question below. Explain your answer.` `Question: [Question]`
		VQA Prompt (ii)	`[VQA Prompt (i)]### Assistant: [LVLM Response]` `### Human: Shorten your answer to the question as much as possible, preferrably only 1 word.`

	A-OKVQA (MC)	VQA Prompt (i)	`### Human: Based on the image, select the correct answer to the question from the options. You MUST mention option labels, i.e., 'A.', 'B.', 'C.' or 'D.' in your response. Explain your answer.`
			`Question: [Question]`
			`Options: A. [Choice 1], B. [Choice 2], C. [Choice 3] , D. [Choice 4]`

		VQA Prompt (ii)	`[VQA Prompt (i)]### Assistant: [LVLM Response]` `### Human: So which option is your final answer: 'A.', 'B.', 'C.' or 'D.'?`
	All	Caption	`### Human: <Img> <ImageHere> </Img>### Human: Describe the image in a couple of sentences.`
	All	Extraction of Details	`### Human: What can you tell me about [entity] in this image?`

LLM	All	Rationale (ii) MiniGPT-4	`### Human: You are given a description of an image, a question and its response below.`
			`Image Content: [Caption Response]`
			`Question: [Question]`
			`Response: [Rationale Response from VQA prompt].`
			`List up to 3 objects or from the image were relevant to answering the question? Describe each object ONLY 2-3 words.### Assistant: Enumerated list of top-3 relevant objects used:`

		Sentence Fusion†	`You are given a question about an image. Modify the question by adding descreptive phrases to entities based on the provided details. Both original and modified questions MUST have similar meaning and answer. [2 Hypothetical Examples]`
			`Question: [Question]`
			`Details: [Bulleted list of entities and 1-2 sentences of corresponding details.]`
			`Modified Question:`

Table 11: All the prompts used in REPARE. '(i)' and '(ii)' indicate a sequential conversation. †For Vicuna model, we add `### Human`, and `### Assistant` prefixes.