

Visual Question Decomposition on Multimodal Large Language Models

Haowei Zhang*¹ Jianzhe Liu*¹ Zhen Han[†]² Shuo Chen³
Bailan He³ Volker Tresp^{3,4} Zhiqiang Xu⁵ Jindong Gu[†]⁶

¹Technical University of Munich, ²Amazon Web Services,

³LMU Munich, ⁴Munich Center for Machine Learning,

⁵MBZUAI, ⁶University of Oxford

{haowei.zhang, jianzhe.liu}@tum.de

 <https://vqd-emnlp2024.github.io/>

Abstract

Question decomposition has emerged as an effective strategy for prompting Large Language Models (LLMs) to answer complex questions. However, while existing methods primarily focus on unimodal language models, the question decomposition capability of Multimodal Large Language Models (MLLMs) has yet to be explored. To this end, this paper explores visual question decomposition on MLLMs. Specifically, we introduce a systematic evaluation framework including a dataset and several evaluation criteria to assess the quality of the decomposed sub-questions, revealing that existing MLLMs struggle to produce high-quality sub-questions. To address this limitation, we propose a specific finetuning dataset, DecoVQA+, for enhancing the model’s question decomposition capability. Aiming at enabling models to perform appropriate selective decomposition, we propose an efficient finetuning pipeline. The finetuning pipeline consists of our proposed dataset and a training objective for selective decomposition. Finetuned MLLMs demonstrate significant improvements in the quality of sub-questions and the policy of selective question decomposition. Additionally, the models also achieve higher accuracy with selective decomposition on VQA benchmark datasets.

1 Introduction

Answering complex questions is a challenging task, especially when the questions require implicit multi-step reasoning to answer. Question Decomposition (QD) is an effective strategy to address this issue. Most related work studies the efficacy of QD with unimodal textual large language models

(LLMs) in enhancing complex textual question answering tasks (Patel et al., 2022; Dua et al., 2022; Zhou et al., 2023; Qi et al., 2023). Although some recent works (You et al., 2023; Qi et al., 2023) have explored question decomposition within the context of visual question answering (VQA) tasks, they follow the paradigm of performing unimodal QD based on the image caption. Typically, they conduct a two-step process: first, generating a caption for the image using a captioning model, and then performing question decomposition using an unimodal textual LLM based on the complex question and the generated image caption. Relying solely on the image caption instead of the image itself may lead to significant information loss.

Recent advancements in Multimodal Large Language Models (MLLMs) have enabled MLLMs to directly perceive image information for answering questions. Yet, how to perform QD on complex visual questions using such MLLMs has been less explored. In the following, we refer to question decomposition using MLLMs on VQA as Visual Question Decomposition (VQD). In this work, we primarily explore the following research questions:

- How can we quantitatively assess the VQD ability of MLLMs? How proficient are existing MLLMs in VQD, or specifically, how is the quality of sub-questions generated by MLLMs?
- How can we enhance the VQD ability of MLLMs and enable the models to properly determine when to decompose and when not to, facing questions with varying difficulties?

To assess the question decomposition capability of MLLMs, a significant obstacle is the absence of metrics for evaluating models’ question decomposition abilities. Recent work (You et al., 2023; Qi et al., 2023) evaluates the model’s question decomposition ability by measuring the final answer’s

* Equal Contribution.

† Corresponding authors:

Zhen Han <hanzhen02111@163.com>, this work does not relate to his position at AWS; Jindong Gu <jindong.gu@outlook.com>

accuracy. However, relying solely on whether a model can correctly answer the original question is an implicit measure of its decomposition ability. Especially, even if the final answer is correct, we have observed various issues in the decomposed sub-questions: for example, some MLLMs produce many repetitive sub-questions, or some sub-questions are entirely irrelevant to the original question, as shown in Figure 1.

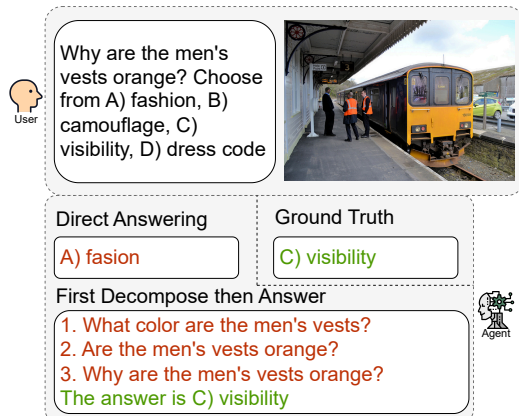


Figure 1: Cases showing that even if the model correctly answers the original question, the generated sub-questions are of low quality: they are irrelevant or repeated from the original question.

Ideally, the sub-questions should be highly relevant to the original question and not repetitive with the original question or other sub-questions. Besides, they should be relatively easy to be grounded, whose answer can be derived from images or pre-trained commonsense knowledge. Figure 2 shows a detailed comparison between sub-questions of high- and low-quality. To this end, we propose SubQuestRater, an evaluation framework for assessing MLLM’s question decomposition ability. Specifically, considering the observed common deficiencies of existing MLLMs’ question decomposition ability, we choose three critical criteria to assess the quality of question decomposition: 1) Non-Repetition, 2) Relevance, and 3) Groundedness. SubQuestRater quantifies the quality of each sub-question by assigning scores based on each criterion.

Besides, it is necessary to have an evaluation dataset containing complex questions requiring decomposition. However, current QA datasets, even specifically aiming at complex reasoning, such as A-OKVQA (Schwenk et al., 2022), still contain a large number of simple questions that do not re-

quire decomposition to answer. Given the lack of publicly available benchmarks solely focusing on complex questions that require decomposition, we introduce a specific question decomposition evaluation dataset. With the help of our proposed evaluation criteria and benchmarks, we evaluate several MLLMs including MiniGPT-v2 (Chen et al., 2023a), LLaVA-1.5 (Liu et al., 2024), etc. The results show that, the decomposed sub-questions generated by these MLLMs cannot perform satisfactorily, demonstrating repetition, irrelevance, or ungroundedness in many cases.

To enhance MLLMs’ capability for VQD, we propose a new finetuning dataset tailored for question decomposition, DecoVQA. It is the first public dataset that consists of manually annotated sub-questions for complex questions. The provided sub-questions feature high quality in the view of non-repetition, relevance and groundedness. To prevent catastrophic forgetting, samples with simple questions in the form of direct answering are added to construct DecoVQA. We finetune MLLMs on this dataset with LoRA (Hu et al., 2021).

Furthermore, we find that existing MLLMs struggle to determine whether they need question decomposition to enhance their reasoning performance when facing problems of varying difficulty. To address this issue, we propose a training pipeline with an upgraded version of DecoVQA, i.e. DecoVQA+, with an extra QA round asking models whether to decompose, and a novel objective function combining a next-token prediction loss (NTP loss) and a binary cross entropy loss (BCE loss) to fine-tune MLLMs. In addition to applying the conventional NTP loss for general reasoning, we design a BCE loss that aims to penalize the errors in deciding whether to decompose questions. Extensive experiments show that MLLMs after finetuning achieve a higher answer accuracy and learn to know when to decompose properly.

To summarize, the main contributions of our work are as follows:

1. We are the first to systematically investigate MLLMs’ ability on visual question decomposition. We propose a comprehensive evaluation framework, SubQuestRater, which includes a benchmark dataset and novel evaluation metrics, to quantitatively evaluate the quality of generated sub-questions from diverse perspectives.
2. We find that existing MLLMs are insufficient to produce sub-questions with high quality. Ef-

efficient finetuning of MLLMs on our proposed dataset, DecoVQA, significantly improves their VQD ability.

3. We propose a finetuning pipeline with an upgraded dataset, DecoVQA+, and a specific training objective for selective VQD, demonstrating improvements in the model’s decision-making regarding decomposing questions or direct answering, as well as the accuracy of final answers.

2 Related Work

2.1 Question Decomposition

Question decomposition has shown impressive capabilities in improving the reasoning performance of language models. Successive Prompting (Dua et al., 2022) and Least-to-Most Prompting (Zhou et al., 2023) are two representative works that break a complicated question into simpler ones iteratively. Decomposed Prompting (Khot et al., 2023) introduces a modular setup of question decomposition, which makes it easy to optimize prompts, pre-trained models and symbolic functions for different sub-tasks. Additionally, question decomposition is capable of increasing the reasoning faithfulness while achieving the accuracy improvement (Radhakrishnan et al., 2023).

Recent studies have explored the potential of VQD. IdealGPT (You et al., 2023) leverages LLMs iteratively to raise sub-questions and determines the final reasoning answer. Socratic Questioning (Qi et al., 2023) utilizes LLMs to generate sub-questions and answer them, stimulating robust recursive thinking. However, all of these studies rely on the reasoning ability of language models, overlooking the visual information that images can bring to question decomposition. The literature most relevant to our work is (Khan et al., 2023), which explored prompting MLLMs to answer VQA questions with question decomposition in the zero-shot settings. However, it only applies VQD as a prompting technique and evaluates whether VQD could enhance the VQA accuracy. That work does not delve into the quality of generated sub-questions along the entire reasoning process, which does not explicitly analyze how well the questions are decomposed.

2.2 Multimodal LLMs

To address the modality gap, MiniGPT-v2 (Chen et al., 2023a) and LLaVA-1.5 (Liu et al., 2024) apply a linear connection layer to connect the

frozen pre-trained vision module and the language model. Besides, they provide a task-oriented instruction training pipeline to decrease instructional ambiguity across various vision-language tasks. GPT-4 with Vision (GPT-4V) (OpenAI et al., 2024) is a powerful MLLM based on instructing GPT-4 and has shown outstanding capability on diverse benchmarks. Besides instruction-following ability, Qwen-VL (Bai et al., 2023b) series outperforms in a range of vision-language tasks, supporting multilingual conversations and dialogues involving multiple interleaved images. Furthermore, InternVL-1.5 (Chen et al., 2024b) introduces a strong vision encoder, dynamic high-resolution images, and a high-quality bilingual dataset to enhance the comprehensive capability of MLLMs further. While these models have demonstrated impressive performance across various benchmarks (Bai et al., 2023b), they continue to struggle with complex tasks that require advanced reasoning (Khan et al., 2023). Numerous techniques, such as parameter-efficient tuning methods like prompting (Gu et al., 2023), in-context learning (Alayrac et al., 2022), and chain-of-thought reasoning (Zhang et al., 2022), can be used to help models handle unseen or complex tasks, but each has its limitations. Parameter-efficient tuning is vulnerable to robustness issues when dealing with out-of-distribution inputs (Chen et al., 2024a), in-context learning often fails to fully utilize multimodal information, focusing predominantly on text (Chen et al., 2023b), and chain-of-thought reasoning is prone to adversarial attacks (Wang et al., 2024). In contrast, our approach employs question decomposition, breaking down complex queries into simpler, more manageable sub-questions, which enhances the models’ reasoning capabilities.

3 How well can MLLMs decompose questions?

Existing works commonly use the accuracy of the final answer to demonstrate a model’s ability to decompose questions. However, this evaluation method is imprecise and implicit. As shown in Figure 1, a MLLM generates sub-questions with low quality, yet can still provide a correct answer. However, these ineffective sub-questions fail to provide the expected assistance in answering the original question and do not help the model’s reasoning process. To address this, we differentiate

the model’s question decomposition skills from the accuracy of answering and propose SubQuestRater, an evaluation framework focusing explicitly on VQD ability. The framework consists of criteria that are specifically designed to assess the quality of sub-questions and an evaluation dataset.

To determine the criteria for the proposed framework, we have analyzed the generated sub-questions by existing MLLMs including MiniGPT-v2, LLaVA, etc., and we have observed the most common issues among them: Some sub-questions repeat the original question or other sub-questions (e.g., semantically equivalent), while others are not relevant to the original question. Besides, some sub-questions cannot be answered from images or commonsense knowledge. These issues largely influence the quality of sub-questions. After conducting the manual review and analysis of the sub-questions, we proposed three criteria to assess the quality of sub-questions, as follows:

Non-Repetition This criterion ensures that sub-questions do not repeat. The definition of repetition here is the case where sub-questions discuss the same topic with the same or different phrasing. For example, in Figure 1, the original question asks why the men’s vests are orange, yet all the sub-questions repeatedly talk about orange vests, causing only redundancy.

Relevance This criterion judges whether a sub-question truly contributes to answering the original question. For example, if the original question asks about the relationship between two people sitting at a table, but the sub-questions inquire about the colors of their clothes or the shapes of the table, these sub-questions are irrelevant. Such distractions can even mislead the model and reduce its performance.

Groundedness This criterion evaluates whether a sub-question can be answered using information directly provided by the image or from commonsense knowledge. Given a relevant and not repeated sub-question, if it can’t be grounded from image or commonsense knowledge, it would still be unhelpful. For example, if the original question asks whether it is safe to cross the road now, and a sub-question inquires what is the time displayed on the traffic light, which helps answer the original question if it can be inferred from the image. However, the image only shows the yellow traffic light and there is no number on the light indicating the

remaining time. Therefore, the sub-question is considered ungrounded.

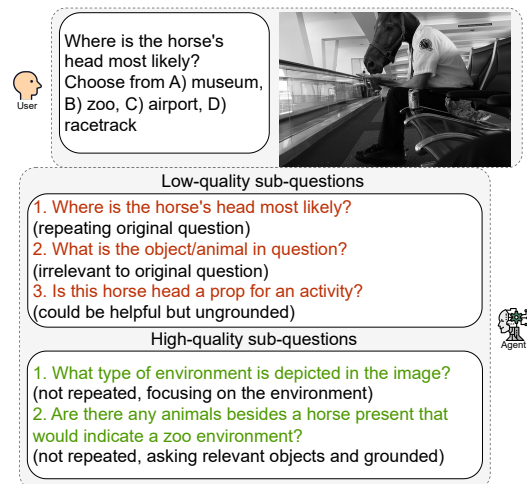


Figure 2: Question decomposition examples of high quality and low quality given a certain image and question.

A more detailed explanation of the criteria with cases is given in Figure 2. By employing this evaluation framework, we have three quantifiable metrics for each sub-question. Algorithm 1 visually demonstrates the complete evaluation process for each sub-question within this framework.

Moreover, we have constructed an evaluation benchmark dataset, since there is currently no dataset composed entirely of complex questions that require decomposition to answer, we construct an evaluation dataset. We manually selected 100 complex questions each from A-OKVQA (Schwenk et al., 2022) and VQA-Introspect (Selvaraju et al., 2020), making a total of 200 questions worth decomposing. A-OKVQA serves as a benchmark necessitating a substantial understanding of external knowledge to formulate accurate responses. VQA-Introspect is a VQA dataset that contains a large number of samples that need complex visual reasoning to answer. A-OKVQA samples are in the form of multiple choice while VQA-Introspect provides open-ended questions. Since these two public datasets have a large number of simple questions which do not need decomposition to answer, we construct an evaluation dataset based on them instead of directly evaluating on them. After establishing the evaluation framework, we choose GPT-4V as the scoring model due to its powerful comprehensive reasoning performance. To ensure that GPT-4V’s judgments align

Criteria	MiniGPT-v2	LLaVA-1.5	Qwen-VL-Chat	InternVL-Chat-V1-5	GPT-4V
Non-Repetition	47.52	42.19	32.10	82.41	97.40
Relevance	36.65	37.33	27.15	73.42	75.36
Groundedness	43.30	44.17	26.49	78.01	84.57

Table 1: Average scores of VQD ability on three criteria of popular existing MLLMs, evaluated with SubQuestRater. The performance of GPT-4V is also provided for reference.

with human judgments, we have conducted alignment experiments. As shown in Appendix A, the results demonstrate that the scoring gap between the judgments of GPT-4V and human beings is small. It is reliable to adopt GPT-4V as the scoring model.

We have measured the VQD ability of popular existing MLLMs with SubQuestRater, including MiniGPT-v2 (Chen et al., 2023a), LLaVA-1.5 (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023b) and InternVL-Chat-V1-5 (Chen et al., 2024b). The results in Table 1 show that these existing models cannot generate satisfactory sub-questions.

4 Enhancing MLLM’s Visual Question Decomposition Capability

Given that the existing MLLMs have poor performance on VQD, this section further explores how to improve the VQD ability of MLLMs. An intuitive method to enhance the decomposition performance of MLLMs is to finetune the models on a dataset tailored for VQD. Specifically, we need a dataset to finetune the models, which exclusively focuses on complex questions with high-quality sub-questions in the view of Non-Repetition, Relevance, and Groundedness. However, there does not exist such a public VQA dataset. Therefore, we propose a specialized dataset, termed DecoVQA, to improve the VQD ability. Furthermore, for effective VQD, models also need to have an improved ability to decide when to decompose questions. We also explain the finetuning pipeline with our proposed dataset and a novel training objective to achieve that goal in detail, as discussed below.

4.1 Dataset Construction of DecoVQA

Question Selection & Decomposition Annotation In our exploration of question decomposition for VQA in Table 1, we recognize that not all questions in existing benchmark datasets necessitate decomposition for answering. Many questions are straightforward and can be addressed without employing a decomposition strategy. Our focus,

therefore, is on questions that demand complex reasoning, making them suitable candidates for the decomposition annotation. For this purpose, we have selected A-OKVQA and VQA-Introspect as our primary data source, as these two datasets contain complex questions requiring external knowledge and visual reasoning to answer respectively.

To identify appropriate samples from A-OKVQA and VQA-Introspect, we adopt specific pre-selection strategies, shown in Appendix F.1 detailedly. After that, we conduct a manual review and pick 200 complex samples that require decomposition from pre-selected samples. Then we manually annotated these samples with logical sub-questions. The details of the annotation process are shown in Appendix F.2.

Dataset Statistics After decomposition annotation, we collected 100 samples from A-OKVQA and 100 samples from VQA-Introspect with high-quality sub-questions from the perspective of Non-Repetition, Relevance, and Groundedness. To prevent overfitting and catastrophic forgetting, we manually picked out another 100 samples from A-OKVQA and 100 samples from VQA-Introspect, which are simple and VQD doesn’t contribute to higher performance for them. These simple samples are added to DecoVQA in the form of direct answering. Overall, DecoVQA has 400 balanced samples in total.

4.2 DecoVQA+

To enhance the capability of MLLMs in selective decomposition, we add an extra QA round on the basis of DecoVQA to enable the models to learn when to decompose properly, facing questions with various difficulties. This extra QA round contains a query asking the models if they would directly answer without any decomposition, given an image and a question. The labels for simple questions are "yes" while the ones for complex questions with human-annotated sub-questions are "no". The extra QA round is added in front of all existing

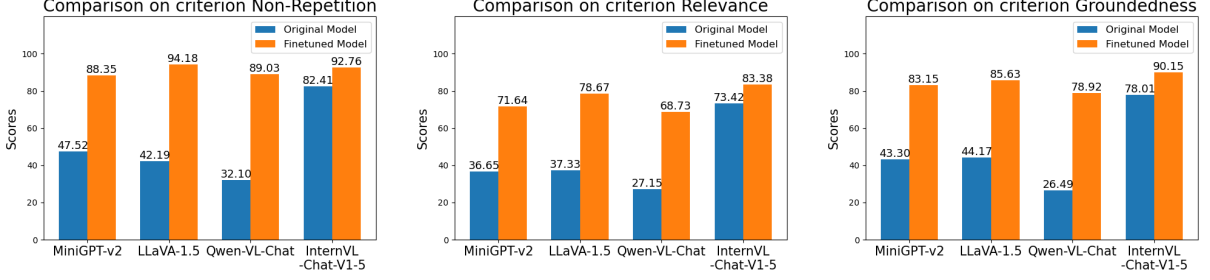


Figure 3: Comparison of VQD ability of different models across three evaluation criteria. Each bar chart represents a specific criterion, comparing the average scores of the original model (in blue) and the corresponding model finetuned with DecoVQA+ (in orange). The vertical axis shows the average scores, while the horizontal axis lists the models. The difference in bar height indicates the performance gain achieved through finetuning.

QA rounds of DecoVQA. We demonstrate the full prompt of a training sample in Figure 6. We refer to this upgraded version as "DecoVQA+".

The superiority of DecoVQA+ is proven in Appendix J, compared to the existing dataset VQA-Introspect. In the ablation study on DecoVQA+, as shown in Table 6, we compare the complete DecoVQA+ to the version with only 100 or 200 samples. The model achieves very similar results after being finetuned with different versions of DecoVQA+. On the one hand, the ablation study proves that our proposed dataset has sufficient samples to train the model. On the other hand, it also indicates that our finetuning pipeline remains efficient, even if there is a lack of high-quality finetuning data in most real-world cases.

To find out whether MLLMs learn to identify questions that need decomposition, we develop an evaluation dataset, Whether2Deco. It consists of 200 simple questions where direct answering is sufficient to answer them correctly and 200 complex questions that need VQD to answer, which are organized in the form of the extra round in DecoVQA+. The questions are equally sampled from A-OKVQA and VQA-Introspect. The statistics of all utilized public datasets and newly proposed datasets are shown in Table 8.

4.3 Training Objective

It is intuitive to finetune MLLMs to improve models' performance on VQD. However, directly applying the conventional next-token prediction loss (NTP loss) on the finetuning for *selective* VQD may not be appropriate. To improve MLLMs' capability of identifying the questions that need to be decomposed, we propose a training objective, SelectiveVQD Loss, combining the NTP loss and a binary cross entropy loss (BCE loss). The BCE

loss aims to penalize the errors in deciding whether to decompose, compared to the labels given in each sample of DecoVQA+.

When the model is asked whether it would perform VQD, we firstly find the specific token position for "yes" or "no" in the sentence, select the logits of these two specific tokens in that position, i.e. "yes" and "no", and then transform these two logits into probabilities through softmax:

$$\begin{aligned}
 \mathbb{P}(yes) &= \mathbb{P}(\hat{w}_s = "yes" | \hat{w}_s \in \{"yes", "no"\}) \\
 &= \text{Softmax}(\text{logit}(\hat{w}_s = "yes"), \text{logit}(\hat{w}_s = "no")) \\
 \mathbb{P}(no) &= \mathbb{P}(\hat{w}_s = "no" | \hat{w}_s \in \{"yes", "no"\}) \\
 &= 1 - \mathbb{P}(yes),
 \end{aligned} \tag{1}$$

where s is the specific token position in the sentence for "yes" or "no" and w_s is the specific token of "yes" or "no".

We compute the BCE loss between these two probabilities, and compute the cumulative NTP loss across all conversation rounds for each sample:

$$BCELoss = -[y_s \log \mathbb{P}(yes) + (1 - y_s) \log(1 - \mathbb{P}(yes))] \tag{2}$$

$$NTPLoss = -\sum_{i=1}^M \log \mathbb{P}(\hat{w}_i = w_i | \hat{w}_{i-1}, \dots, \hat{w}_1), \tag{3}$$

where y_s is the binary label indicating whether a specific sample needs decomposition or not. w_i denotes the i -th token in the ground truth sentence while \hat{w}_i denotes the predicted i -th token and M is the number of tokens of the prediction.

The final combined SelectiveVQD Loss is a weighted sum of both NTP loss and BCE loss:

$$SelectiveVQDLoss = \sum_{j=1}^N (\lambda \cdot NTPLoss_j + \omega \cdot BCELoss_j), \tag{4}$$

where j denotes the j -th training sample and N denotes the total number of training samples. The NTP loss is computed for the entire training sample,

and the BCE loss focuses specifically on determining whether to decompose the given question in the selective stage. λ and ω are two tunable hyperparameters to balance the weights of the two losses in the final combined loss.

5 Experiments

5.1 Experiment Setup

Models With SubQuestRater, we compare the VQD ability of four popular MLLMs: MiniGPT-v2, LLaVA-1.5, Qwen-VL-Chat, and InternVL-Chat-V1-5 before and after finetuning on our proposed datasets. We finetune all these MLLMs on DecoVQA, DecoVQA+, and DecoVQA+ with SelectiveVQD Loss to see the improvement in their VQD capability. Additionally, we evaluate the improvement in VQA accuracy and the models' capability to appropriately determine when to decompose questions through finetuning.

Datasets The finetuning datasets include DecoVQA and DecoVQA+. DecoVQA+ adds an extra QA round based on DecoVQA, asking models whether to decompose questions before decomposition. As for used evaluation datasets, we assess the VQD capability of MLLMs before and after finetuning on the proposed evaluation dataset in SubQuestRater, which contains 200 complex questions. The prompt for evaluating the VQD ability is shown in Figure 5. Besides, we evaluate the VQA accuracy on A-OKVQA, GQA (Hudson and Manning, 2019), and VQA-Introspect, containing complex reasoning questions (please refer to Appendix D.1 for more statistical details). As for A-OKVQA and VQA-Introspect, the subsets of data used for inference are different from the subsets selected for constructing our finetuning datasets, preventing the problem of data leakage. We also evaluate the accuracy on Whether2Deco to test whether MLLMs are able to determine when to decompose questions properly. The prompt for evaluation experiments on accuracy is under selective VQD setting, which is in the same form of samples in DecoVQA+ shown in Figure 6.

5.2 Quantitative Evaluation

The quantitative evaluation involves two parts: *evaluation of decomposed sub-questions* under SubQuestRater framework and *accuracy comparison* on VQA datasets and Whether2Deco. The finetuning is efficient based on the dataset with a small

number of samples. The supplementary evaluation on MMBench (Liu et al., 2023) in Appendix I shows our finetuning does not hurt the all-around performance of MLLMs, but even slightly improves the comprehensive performance in many aspects.

Evaluation of Decomposed Sub-questions To compare the VQD ability of the MLLMs before and after finetuning, we conduct evaluation with the SubQuestRater framework. Figure 3 illustrates the comparison of average scores of sub-questions generated by four MLLMs before and after finetuning with DecoVQA+. It can be observed that finetuned models have outperformed their original versions on all three criteria, indicating the VQD ability has been enhanced significantly through finetuning. Some of the models even show nearly double the scores in some criteria after finetuning. In addition to the average score, we also compare the number of samples before and after finetuning, which achieve a high score (75-100) and a low score (0-25) on three criteria, as shown in Figure 8. The results show that there are considerably more high-scored samples and less low-scored samples through finetuning. The VQD abilities of other finetuned checkpoints are listed in Table 9. It shows that all finetuned versions of MLLMs have an improvement in VQD ability compared to the original model.

Additionally, we have conducted experiments by varying the number of samples in DecoVQA, which ultimately lead to similar results, as shown in Table 5, indicating that there is no need to add more samples to DecoVQA+ for further improvements.

VQA Accuracy & Whether2Deco Accuracy Firstly, we investigate whether better VQD performance leads to higher accuracy. To prove that point, we compare the accuracy of models finetuned with DecoVQA and their corresponding baselines, as shown in the second and the first row of each model respectively in Table 2. It is clear that the models with higher VQD capability through finetuning achieve higher accuracy than their original versions in most experiments.

Existing MLLMs are unable to decide when to decompose appropriately and tend to make a fifty-fifty guess, as shown in the first line of each model. For improving the performance in selective decomposition, it is very important for models to learn when to decompose, facing questions with varying difficulties, since unnecessary decomposition may

Models	A-OKVQA	GQA	VQA-Introspect	Whether2Deco
MiniGPT-v2	41.2	44.2	62.1	46.8
finetuned by DecoVQA	60.6 \uparrow (+19.4)	50.4 \uparrow (+6.2)	71.8 \uparrow (+9.7)	42.8 \downarrow (-4.0)
finetuned by DecoVQA+	60.7 \uparrow (+19.5)	50.7 \uparrow (+6.5)	72.1 \uparrow (+10.0)	61.0 \uparrow (+14.2)
finetuned by DecoVQA+ with SelectiveVQD Loss	64.0 \uparrow (+22.8)	51.7 \uparrow (+7.5)	72.5 \uparrow (+10.4)	71.5 \uparrow (+24.7)
LLaVA-1.5	67.7	52.1	67.2	49.3
finetuned by DecoVQA	69.4 \uparrow (+1.7)	52.8 \uparrow (+0.7)	73.5 \uparrow (+6.3)	4.8* \downarrow (-44.5)
finetuned by DecoVQA+	72.7 \uparrow (+5.0)	57.2 \uparrow (+5.1)	75.4 \uparrow (+8.2)	68.8 \uparrow (+19.5)
finetuned by DecoVQA+ with SelectiveVQD Loss	73.9 \uparrow (+6.2)	56.7 \uparrow (+4.6)	75.8 \uparrow (+8.6)	75.0 \uparrow (+25.7)
Qwen-VL-Chat	71.4	53.5	77.8	48.0
finetuned by DecoVQA	72.0 \uparrow (+0.6)	58.0 \uparrow (+4.5)	75.9 \downarrow (-1.9)	43.3 \downarrow (-4.7)
finetuned by DecoVQA+	73.1 \uparrow (+1.7)	59.3 \uparrow (+5.8)	83.6 \uparrow (+5.8)	58.8 \uparrow (+10.8)
finetuned by DecoVQA+ with SelectiveVQD Loss	73.3 \uparrow (+1.9)	59.1 \uparrow (+5.6)	83.9 \uparrow (+6.1)	61.8 \uparrow (+13.8)
InternVL-Chat-V1-5	80.7	64.8	80.5	58.3
finetuned by DecoVQA	83.5 \uparrow (+2.8)	66.4 \uparrow (+1.6)	86.0 \uparrow (+5.5)	53.5 \downarrow (-4.8)
finetuned by DecoVQA+	83.3 \uparrow (+2.6)	66.5 \uparrow (+1.7)	86.9 \uparrow (+6.4)	67.0 \uparrow (+8.7)
finetuned by DecoVQA+ with SelectiveVQD Loss	83.7 \uparrow (+3.0)	66.8 \uparrow (+2.0)	87.3 \uparrow (+6.8)	68.3 \uparrow (+10.0)

Table 2: Comparison of VQA accuracy (%) on external knowledge (A-OKVQA) and visual reasoning (GQA and VQA-Introspect) datasets and Whether2Deco accuracy (%) before and after fine-tuning MLLMs. DecoVQA+ is constructed based on DecoVQA, with an extra QA round asking MLLMs whether the question needs VQD to answer or not. *Here LLaVA-1.5 fails to follow the pre-defined answering template.

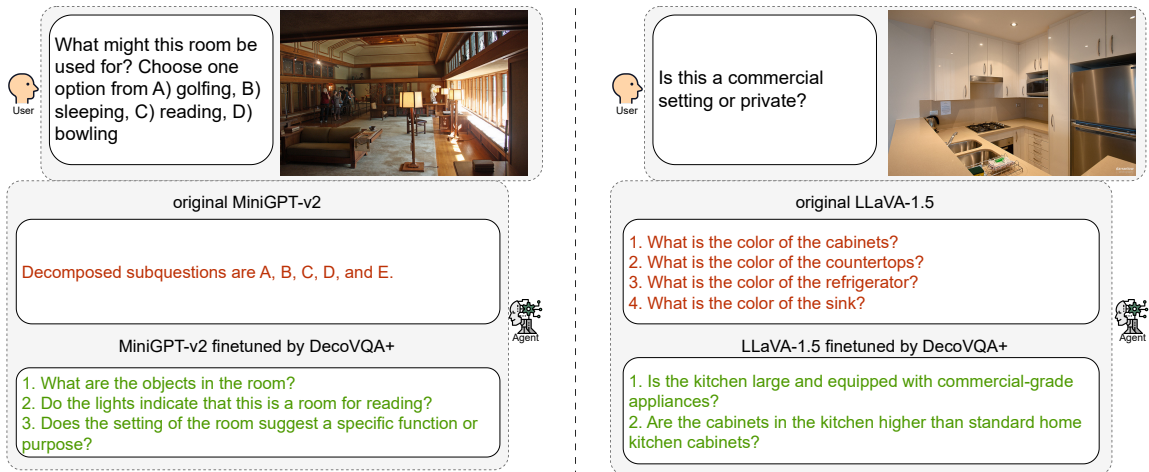


Figure 4: Cases showing the comparison of question decomposition by different models before and after finetuning. The left image demonstrates MiniGPT-v2’s decomposition on A-OKVQA, while the right image shows LLaVA-1.5’s decomposition on VQA-Introspect.

mislead the reasoning process. From the second row of each model, we can see that higher quality of sub-questions does not mean better performance on determining when to decompose questions. To address this problem, we finetune the models with DecoVQA+. As shown in the third row of each model in Table 2, there is a significant improvement in the accuracy on Whether2Deco after fine-tuning with DecoVQA+, compared to the baseline and the checkpoint finetuned by DecoVQA. Moreover, the accuracy of VQA tasks also increases because of the better whether-to-decompose policy of the

finetuned models.

To further enhance the ability of MLLMs to perform selective decomposition, we train the models with SelectiveVQD Loss. In contrast to the training with only the NTP loss, the models achieve higher accuracy on Whether2Deco and also VQA tasks in most cases. If compared to the original models before finetuning, the accuracy on all evaluation datasets increases significantly. The results of evaluation experiments with different random seeds in Figure 9 show the stable effectiveness of our entire pipeline. We also compare our proposed VQD

pipeline with the existing paradigm of unimodal QD based on the image caption, as shown in 12. The results demonstrate that VQD outperforms the unimodal QD method. The comparison between our finetuning and In-context Learning proposed in (Khan et al., 2023) is shown in Appendix L.

5.3 Qualitative Evaluation

In this subsection, we will use several examples to visually illustrate the changes of sub-questions before and after finetuning with DecoVQA+. Figure 4 shows that the quality of the sub-questions has indeed been significantly improved after finetuning. The sub-questions generated by finetuned models are not repetitive, relevant to the original question and grounded, instead of ineffective decomposition or low-quality sub-questions originally. More case studies are shown in Figure 15.

6 Conclusion

This paper systematically investigates VQD capabilities on MLLMs. We propose a systematic evaluation framework for VQD, SubQuestRater, including a dataset and evaluation metrics to quantitatively measure the generated sub-questions by MLLMs. SubQuestRater is applied to popular MLLMs and we find that they are inadequate to produce high-quality sub-questions. To enhance the capability of MLLMs to decompose questions, a specialized dataset DecoVQA with human-annotated sub-questions is proposed. To further improve the ability to perform selective VQD, we propose a training pipeline with an upgraded dataset DecoVQA+ and a novel training objective. Finetuned MLLMs demonstrate significant improvement in the quality of generated sub-questions and the policy of whether-to-decompose. Additionally, the models also achieve higher VQA accuracy under selective VQD through finetuning on our proposed datasets.

Limitations

The main limitations in our work include: 1) Question Decomposition can be extended into complex task decomposition for an agent (multiple sub-tasks), leaving it as future work. 2) We apply finetuning to increase MLLM’s VQD ability, which requires the model’s detailed parameter information. Thus, community users could not apply our method for enhancing closed-source MLLMs.

Acknowledgement

The authors acknowledge support by the German Federal Ministry for Education and Research (BMBF), funding project Software Campus 2.0 / C-R-KG (FKZ 01IS17048).

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). *Preprint*, arXiv:2204.14198.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Jinze Bai, Shuai Bai, Yunfei Chu, and et al. 2023a. [Qwen technical report](#). *Preprint*, arXiv:2309.16609.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Zheng Cai, Maosong Cao, Haojiong Chen, and et al. 2024. [Internlm2 technical report](#). *Preprint*, arXiv:2403.17297.
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. [Minigt-v2: large language model as a unified interface for vision-language multi-task learning](#). *Preprint*, arXiv:2310.09478.
- Shuo Chen, Jindong Gu, Zhen Han, Yunpu Ma, Philip Torr, and Volker Tresp. 2024a. Benchmarking robustness of adaptation methods on pre-trained vision-language models. *Advances in Neural Information Processing Systems*, 36.
- Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. 2023b. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 1(2).
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hwei Guo, Conghui He,

- Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. [How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites](#). *Preprint*, arXiv:2404.16821.
- 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. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. [Successive prompting for decomposing complex questions](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics (international student edition)*. Pisani, R. Purves, 4th edn. WW Norton & Company, New York.
- Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. 2023. [A systematic survey of prompt engineering on vision-language foundation models](#). *arXiv preprint arXiv:2307.12980*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Drew A. Hudson and Christopher D. Manning. 2019. [Gqa: A new dataset for real-world visual reasoning and compositional question answering](#). *Preprint*, arXiv:1902.09506.
- Zaid Khan, Vijay Kumar B G, Samuel Schuler, Manmohan Chandraker, and Yun Fu. 2023. [Exploring question decomposition for zero-shot vqa](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 56615–56627. Curran Associates, Inc.
- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. [Decomposed prompting: A modular approach for solving complex tasks](#). *Preprint*, arXiv:2210.02406.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). *Preprint*, arXiv:2310.03744.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. [Mmbench: Is your multi-modal model an all-around player?](#) *Preprint*, arXiv:2307.06281.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. [Ok-vqa: A visual question answering benchmark requiring external knowledge](#). *Preprint*, arXiv:1906.00067.
- OpenAI, Josh Achiam, Steven Adler, and et al. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pruthvi Patel, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. [Is a question decomposition unit all we need?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4569, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. [The art of SOCRATIC QUESTIONING: Recursive thinking with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.
- Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. [Question decomposition improves the faithfulness of model-generated reasoning](#). *Preprint*, arXiv:2307.11768.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). *Preprint*, arXiv:1908.10084.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. [A-okvqa: A benchmark for visual question answering using world knowledge](#). *Preprint*, arXiv:2206.01718.
- Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. [Squinting at vqa models: Introspecting vqa models with sub-questions](#). In *CVPR 2020*.
- Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. 2024. [Stop reasoning! when multimodal](#)

llms with chain-of-thought reasoning meets adversarial images. *Conference on Language Modeling*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. *IdealGPT: Iteratively decomposing vision and language reasoning via large language models*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11289–11303, Singapore. Association for Computational Linguistics.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of Biostatistics*, 7.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. *Least-to-most prompting enables complex reasoning in large language models*. *Preprint*, arXiv:2205.10625.

A Alignment of Judgements from GPT-4V and Human Reviewers

To ensure that the judgments from GPT-4V and human beings are highly aligned, human reviewers manually evaluate the sub-questions generated by MiniGPT-v2 and its finetuned checkpoint on three criteria defined in SubQuestRater. We compare the judgments from GPT-4V and human reviewers from three perspectives: 1) Comparison of average score, 2) Pearson correlation coefficient (Freedman et al., 2007), 3) Spearman correlation coefficient (Zar, 2005). The comparison of average score serves as a coarse-grained evaluation of overall alignment, with results shown in Table 3. Pearson and Spearman correlation coefficients assess linear and monotonic relationships between two sets of judgments respectively, with results shown in Table 4. The results demonstrate that the judgments from GPT-4V and human reviewers on all three criteria are highly aligned.

B Ablation Studies

The results of ablation studies, as shown in Table 5 and in Table 6. The samples from all the versions of DecoVQA+ follow a balanced distribution. For instance, DecoVQA+100 has 25 complex MC and 25 complex open-ended questions that need VQD while it has the other 25 simple MC and 25 simple open-ended questions that do not need VQD to answer. The models finetuned by DecoVQA+ with

Model	Criteria	GPT-4V	Human	Error Rate
Original MiniGPT-v2	Non-Repetition	51.92	52.37	-0.86%
	Relevance	40.86	39.34	3.86%
	Groundedness	47.65	47.84	-0.40%
MiniGPT-v2 finetuned by DecoVQA	Non-Repetition	94.22	93.79	0.46%
	Relevance	74.54	75.79	-1.65%
	Groundedness	86.36	87.47	-1.27%

Table 3: Comparison of average scores of the judgements on SubQuestRater dataset from GPT-4V and human reviewers on three criteria. We regard judgements from human reviewers as the ground truth when computing the error rate.

Model	Criteria	Pearson	Spearman
Original MiniGPT-v2	Non-Repetition	0.828	0.820
	Relevance	0.813	0.813
	Groundedness	0.801	0.795
MiniGPT-v2 finetuned by DecoVQA	Non-Repetition	0.867	0.864
	Relevance	0.804	0.784
	Groundedness	0.832	0.886

Table 4: Pearson and Spearman correlation coefficients of the judgements on SubQuestRater dataset from GPT-4V and human reviewers on three criteria. All results are statistically highly significant (p-value < 0.001).

varying sample numbers generate similar results on VQD ability, VQA accuracy, and Whether2Deco accuracy, demonstrating that 400 samples are sufficient to ensure both efficient and reliable finetuning.

MiniGPT-v2	DecoVQA+100	DecoVQA+200	DecoVQA+400
Non-Repetition	91.56	91.48	88.35
Relevance	70.19	72.59	71.64
Groundedness	87.35	85.49	83.15
LLaVA-1.5	DecoVQA+100	DecoVQA+200	DecoVQA+400
Non-Repetition	71.92	88.97	94.18
Relevance	67.92	79.41	78.67
Groundedness	80.94	89.28	85.63

Table 5: Ablation study about VQD ability on finetuning models with DecoVQA+ with a varying sample number. DecoVQA+400 is the version of DecoVQA+ with which we finetune MLLMs in other experiments.

C Experiment Details

C.1 Models

The versions of models we use are listed as follows, corresponding official tokenizers are applied for all the models:

- **MiniGPT-v2**, which is based on Llama2-Chat-7B-HF (Touvron et al., 2023).

Models	A-OKVQA	GQA	VQA-Introspect	Whether2Deco
MiniGPT-v2	41.2	44.2	62.1	46.8
finetuned by DecoVQA+100	59.3 ↑ (+18.1)	51.4 ↑ (+7.2)	70.1 ↑ (+8.0)	54.3 ↑ (+7.5)
finetuned by DecoVQA+200	61.3 ↑ (+20.1)	50.9 ↑ (+6.7)	73.4 ↑ (+11.3)	63.8 ↑ (+17.0)
finetuned by DecoVQA+400	60.7 ↑ (+19.5)	50.7 ↑ (+6.5)	72.1 ↑ (+10.0)	61.0 ↑ (+14.2)
LLaVA-1.5	67.7	52.1	67.2	49.3
finetuned by DecoVQA+100	73.4 ↑ (+5.7)	53.6 ↑ (+1.5)	72.5 ↑ (+5.3)	56.0 ↑ (+6.7)
finetuned by DecoVQA+200	74.4 ↑ (+6.7)	57.9 ↑ (+5.8)	78.7 ↑ (+11.5)	69.0 ↑ (+19.7)
finetuned by DecoVQA+400	72.7 ↑ (+5.0)	57.2 ↑ (+5.1)	75.4 ↑ (+8.2)	68.8 ↑ (+19.5)

Table 6: Ablation study about VQA accuracy and Whether2Deco accuracy on finetuning models with DecoVQA+ with a varying sample number. DecoVQA+400 is the version of DecoVQA+ with which we finetune MLLMs in other experiments.

- **LLaVA-1.5**, which is based on Vicuna-13B v1.5 (Chiang et al., 2023) and with lora as the pretraining schedule.
- **Qwen-VL-Chat**, which is based on Qwen-7B (Bai et al., 2023a).
- **InternVL-Chat-V1-5**, which is based on InternLM2-20B (Cai et al., 2024).
- **GPT-4-vision-preview**, which is based on GPT-4 (OpenAI et al., 2024).

C.2 Finetuning Settings

For all the mentioned open-source MLLMs, we use their official GitHub repository code to perform LoRA finetuning on the connection layer between the two modalities. All of the models are trained on $2 \times$ A40 GPU until the training loss converges.

C.3 Inference Settings

We use a batch size of 1 in all inference tasks. Greedy search is used for all inferences. The parameters may be sub-optimal.

C.4 Prompts

C.4.1 Prompt for Scoring VQD Ability

The complete prompt for scoring VQD ability, or specifically, the quality of sub-questions is shown in Figure 5.

C.4.2 Prompt for Selective VQD

As shown in Figure 6, firstly, we perform a selective stage, which asks the model whether to decompose the question. If the model answers that it can directly answer the question without decomposition by "Yes", then implement the direct answering; if the model answers that it needs to decompose the question firstly by "No", then implement a three-phase decomposition process.

D Datasets

D.1 Public Datasets

The statistics of used public datasets are listed in Table 7.

- **A-OKVQA** (Schwenk et al., 2022) is a complex knowledge-based benchmark for VQA. As an augmented version of OK-VQA (Marino et al., 2019), the questions in A-OKVQA are not only diverse but also require a wide-ranging common-sense and knowledge outside the image to answer. A-OKVQA has both multiple choice and open-ended question forms for each sample, and we select multiple choice here to cover more question types in the experiments.
- **GQA** (Hudson and Manning, 2019) features compositional questions related to real-world images, utilizing semantic representations of both scenes and questions to reduce language priors and conditional influences.
- **VQA-Introspect** (Selvaraju et al., 2020) is a new dataset based on a reasoning split from VQA (Antol et al., 2015) dataset, which contains complex reasoning questions with the open-ended form. VQA-Introspect consists of 200K perception questions as sub-questions to help answer difficult reasoning questions. Though this public dataset provides us a large number of sub-questions, questions with varying difficulties are mixed together and no label is pointing it out, which leads to bad finetuning results for selective decomposition. We randomly sampled 3,000 questions for the evaluation on VQA-Introspect.

D.2 Proposed Datasets

The statistics of proposed datasets are listed in Table 8.

E Quantitative Evaluation for VQD Ability

F Details of Data Construction

F.1 Pre-selection Strategies for Selecting Samples


To identify questions from the A-OKVQA dataset that would benefit from decomposition, we employ a specific pre-selection strategy. Initially, we used a MLLM to perform zero-shot inference on the dataset, a process we term "direct inference". Subsequently, we engaged the same model in another

User


You have 3 tasks:
 Evaluate the following texts based on the given image.
 If a sub-question is irrelevant to the main question and does not help in answering it at all (e.g., the main question is asking about relationship between two person sitting at the table, but the sub-questions are meaninglessly asking about colors of their clothes, or shapes of the table), classify it B. Otherwise, classify it G.
 If a sub-question is a repetition of the main question or any existing sub-questions (repetition means the sub-question repeats exactly the same content or discusses the same topic in a different form), classify it R. Otherwise, classify it U.
 If the answer to a sub-question can be derived from the image through direct observation, basic knowledge, logical inference, or reasonable assumptions, classify it Y. Otherwise, if the sub-question requires information that is not available in the image, classify it N.
 In conclusion, I want 3 classes for each sub-question, G/B, Y/N, U/R.
Attention:
 1. Do not repeat the sub-question or give explanation, just give me the 3 classes.
 2. If you can not find any sub-question under the answer of a main question, cases can be either the answers are not presented in a sub-question form or the sub-question is incomplete, classify it E.

Here is the main question:
 What service does the red bus connect passengers to? Choose from A) subway service, B) tram service, C) train service, D) plane service.

Here are the sub-questions:
 1. What transportation mode does the bus connect passengers to?
 2. Does the presence of a car in the image indicate that the airport is nearby?



1. G, Y, R
 2. B, N, U


Agent


(a) An example of evaluation on effective sub-questions.

User


You have 3 tasks:
 Evaluate the following texts based on the given image.
 If a sub-question is irrelevant to the original question and does not help in answering it at all (e.g., the original question is asking about relationship between two person sitting at the table, but the sub-questions are meaninglessly asking about colors of their clothes, or shapes of the table), classify it B. Otherwise, classify it G.
 If a sub-question is a repetition of the original question or any existing sub-questions (repetition means the sub-question repeats exactly the same content or discusses the same topic in a different form), classify it R. Otherwise, classify it U.
 If the answer to a sub-question can be derived from the image through direct observation, basic knowledge, logical inference, or reasonable assumptions, classify it Y. Otherwise, if the sub-question requires information that is not available in the image, classify it N.
 In conclusion, I want 3 classes for each sub-question, G/B, Y/N, U/R. **Attention:**
 1. Do not repeat the sub-question or give explanation, just give me the 3 classes.
 2. If you can not find any sub-question under the answer of a original question, cases can be either the answers are not presented in a sub-question form or the sub-question is incomplete, classify it E.

Here is the original question:
 In which manner were the desserts here prepared? Choose from A) baking, B) open fire, C) grilling, D) frying.

Here are the sub-questions:
 A, B, C and D



E


Agent

(b) An example of evaluation on ineffective sub-questions (error).

Figure 5: Prompt for scoring the quality of sub-questions with GPT-4V.

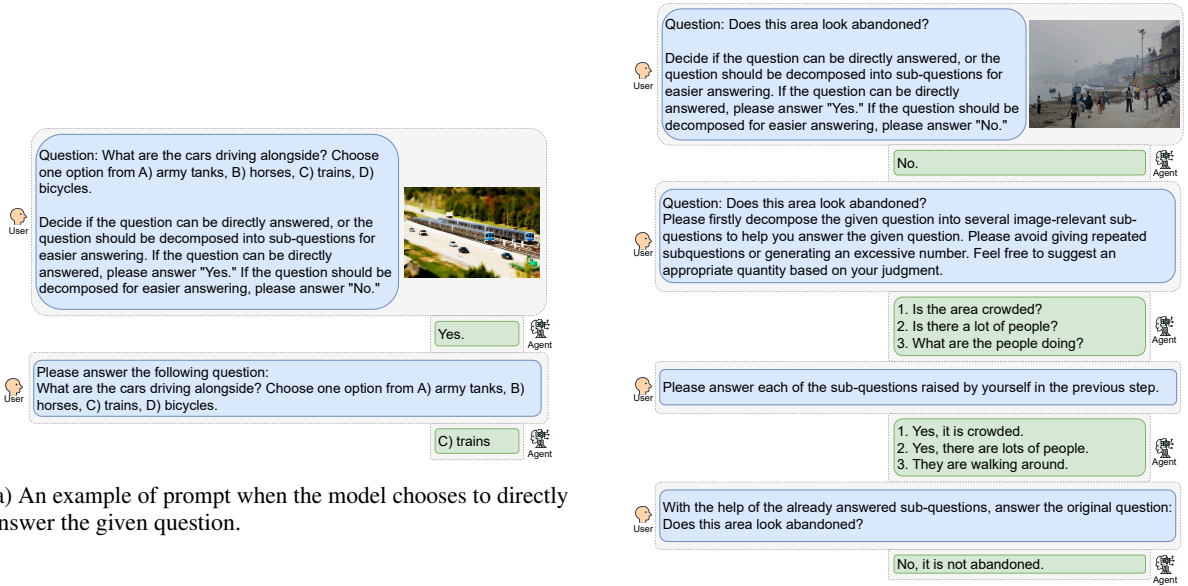


Figure 6: Prompt of selective decomposition samples in DecoVQA+.

Dataset	Dataset Type	Question Type	# Images	# Questions
A-OKVQA	external knowledge	multiple choice	6,030	6,702
GQA	visual reasoning	open-ended questions	398	12,578
VQA-Introspect	visual reasoning	open-ended questions	17,495	22,793

Table 7: Experimental statistics for public datasets used in the paper.

Dataset	Usage	Motivation	# Images	# Questions
SubQuestRater	evaluation	measuring the quality of sub-questions	200	200
DecoVQA	finetuning	improving the VQD ability	397*	400
DecoVQA+	finetuning	improving the selective VQD ability	397*	400
Whether2Deco	evaluation	testing the models' ability to identify whether a question requires decomposition	395*	400

Table 8: Experimental statistics for proposed datasets in the paper. *Several images correspond to more than one question.

Algorithm 1: Evaluation algorithm for the quality of sub-questions

q : Sub-question
 \mathbb{Q} : Set of sub-questions from one sample
 b_1, b_2, b_3 : Binary score for 3 criteria of a sub-question
 B_1, B_2, B_3 : Lists of binary scores for 3 criteria of sub-questions
 s_1, s_2, s_3 : Score for 3 criteria for a sample
Check if there are effective sub-questions
if $\mathbb{Q} == \emptyset$ **then**
 $s_1 = 0$
 $s_2 = 0$
 $s_3 = 0$
else
 for q **in** \mathbb{Q} **do**
 $b_1, b_2, b_3 = \{ScoreModel(q) \mid q \in \mathbb{Q}\}$
 for $i \in [1,3]$ **do**
 AppendtoList(b_i, B_i)
 for $j \in [1,3]$ **do**
 $s_i = CalculateAverage(B_j)$
return s_1, s_2, s_3

round of zero-shot inference, but this time utilizing a question decomposition prompt. In this round, the model is asked to decompose the main question into sub-questions, then answer these, and finally proceed to answer the main question. We refer to this method as "decompose inference". We choose MiniGPT-v2 as the multimodal LLM here. Our primary focus was on questions that were incorrectly answered in direct inference but correctly in decompose-inference, as these exhibited a high likelihood of requiring decomposition.

To find appropriate samples from the VQA-Introspect dataset, we adopt an automated pre-selection strategy based on BLEU (Papineni et al., 2002) metric. BLEU metric is originally used to measure the quality of machine translation, here we use it as a metric to assess repetition. Since VQA-Introspect has provided a large number of redundant sub-questions, we firstly filter out semantically repetitive sub-questions for each sample to prevent from overfitting. A higher BLEU score between two sub-questions means that one of the sub-questions is repetitive. Then we set a threshold number and choose the samples with a remaining number of sub-questions exceeding the threshold.

F.2 Annotation Process

Given the proficiency of GPT-4V in VQD, as shown in Table 1, we utilize GPT-4V to generate initial sets of decomposed sub-questions for each selected sample. Subsequently, we perform a meticulous manual review to these sub-questions. During this process, we eliminate sub-questions that do not contribute meaningfully towards answering the main question and also remove redundant sub-questions that share similar semantic content. Additionally, we supplement the sets with new sub-questions in instances where the decomposition logic appears incomplete, ensuring a more thorough and effective decomposition process.

G Robust Evaluation for MC Datasets

To compute the accuracy of the inference results under multiple choice setting, since an exact match of either option index or word can lead to serious underestimation, the first step is to map the model answer into one of four options. We have designed a robust algorithm to evaluate the accuracy of multiple choice based on the method provided by A-OKVQA. As demonstrated in Alogrithm 2, if no or several options are detected in the model

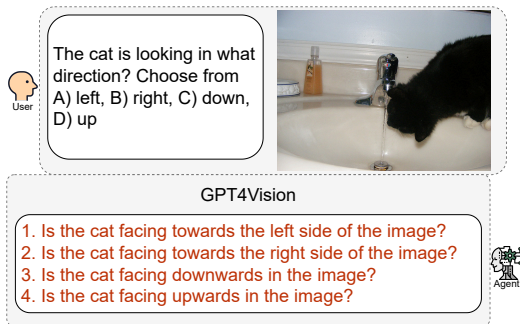


Figure 7: In some cases, GPT-4V will also produce sub-questions that do not fit our criteria.

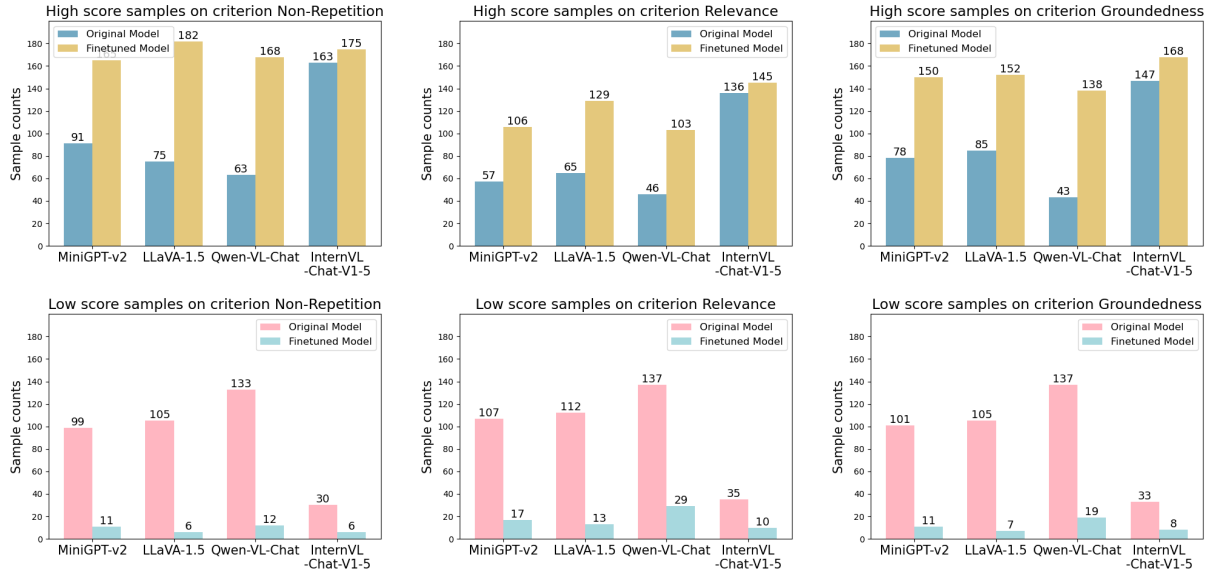


Figure 8: Comparison of VQD ability of different models across three evaluation criteria. Each bar chart represents a specific criterion. The first row compares the number of the high-scored (75-100) samples generated by the original model (in cyan) and the corresponding model finetuned with DecoVQA+ (in yellow). The second row compares the number of the low-scored (0-25) samples generated by the original model (in pink) and the corresponding model finetuned with DecoVQA+ (in blue). The vertical axis shows the number of high-scored samples or low-scored samples, while the horizontal axis lists the models. The difference in bar height indicates the performance gain achieved through finetuning.

MiniGPT-v2	original Model	finetuned by DecoVQA	finetuned by DecoVQA+	finetuned by DecoVQA+ with SelectiveVQD Loss
Non-Repetition	47.52	93.72	88.35	90.58
Relevance	36.65	74.17	71.64	73.73
Groundedness	43.30	85.98	83.15	84.53
LLaVA-1.5	original Model	finetuned by DecoVQA	finetuned by DecoVQA+	finetuned by DecoVQA+ with SelectiveVQD Loss
Non-Repetition	42.19	92.04	94.18	92.68
Relevance	37.33	81.62	78.67	78.48
Groundedness	44.17	86.19	85.63	84.39
Qwen-VL-Chat	original Model	finetuned by DecoVQA	finetuned by DecoVQA+	finetuned by DecoVQA+ with SelectiveVQD Loss
Non-Repetition	32.10	80.66	89.03	89.15
Relevance	27.15	69.52	68.73	67.15
Groundedness	26.49	77.34	78.92	77.51
InternVL-Chat-V1-5	original Model	finetuned by DecoVQA	finetuned by DecoVQA+	finetuned by DecoVQA+ with SelectiveVQD Loss
Non-Repetition	82.41	87.40	92.76	94.11
Relevance	73.42	81.11	83.38	83.30
Groundedness	78.01	87.62	90.15	89.47

Table 9: Comparison of VQD abilities of all the original models and their corresponding finetuned versions.

answer during the exact match step, we use SentenceTransformer (Reimers and Gurevych, 2019) to map the model answer to one option or use GPT-4 to do the mapping when the answer is too long. We have observed that if the answer sentence is too long, especially when there is more than one option mentioned in the answer, the mapping by SentenceTransformer tends to be random and misleading. For computing the accuracy over open-ended questions, given that the reference answer in used datasets has only one or two words, if the reference answer is mentioned in the model output, the output is considered as correct.

Algorithm 2: Robust algorithm for measuring accuracy on MC datasets

a : Model answer
 \hat{a} : Mapped option
 n : Number of exact match options
 τ : Threshold of sentence length
Attempt exact match and get n mentioned options from model answer
if $n == 1$ **then**
 | $\hat{a} = \text{ExactMatch}(a)$
else
 | **if** $\text{len}(\text{tokenize}(a)) \leq \tau$ **then**
 | $\hat{a} = \text{SentenceTransformer}(a)$
 | **else**
 | $\hat{a} = \text{GPT-4}(a)$
return \hat{a}

H Variance

To verify the stability of our proposed method, each experiment was done with three different random seeds, while keeping other settings unchanged. The variance results in Figure 9 show that random seeds influence the accuracy of the model output very slightly.

I Does the finetuning hurt the all-around performance?

Finetuning may lead to catastrophic forgetting, which hurts the essential all-around performance of MLLMs. MMBench (Liu et al., 2023) is a systematic pipeline that evaluates the comprehensive abilities of MLLMs. Figure 10, 11, 12 and 13 demonstrate the evaluation results of different checkpoints on MMBench. It shows that our finetuning does not do harm to most of the abilities, while some of them are even improved after finetuning.

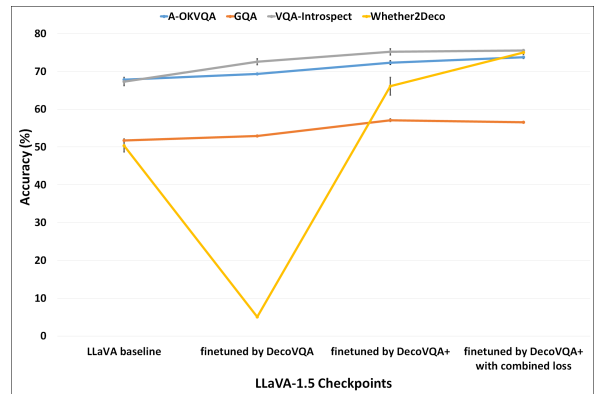
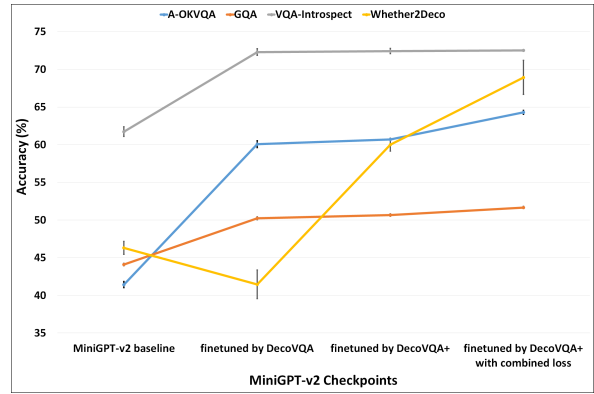


Figure 9: Variance of inference experiments with MiniGPT-v2 and LLaVA-1.5, plotted as error bars. Each experiment is conducted with three different random seeds, keeping other settings unchanged.

J Finetuning with DecoVQA+ vs. with VQA-Introspect

The existing public dataset VQA-Introspect has already provided us with complex visual reasoning questions with sub-questions. However, not all questions are complex enough to require decomposition, and a large number of provided sub-questions are repetitive and superficial. To compare with the quality of our proposed dataset, we also finetune MLLMs with the entire training set of VQA-Introspect (excluding the samples used in the evaluation experiments). As shown in Table 10 and 11, the performance of the MLLMs finetuned with DecoVQA+ is much better than the ones finetuned with VQA-Introspect. The results demonstrate that the quality of our proposed dataset outperforms the existing public dataset with sub-questions.

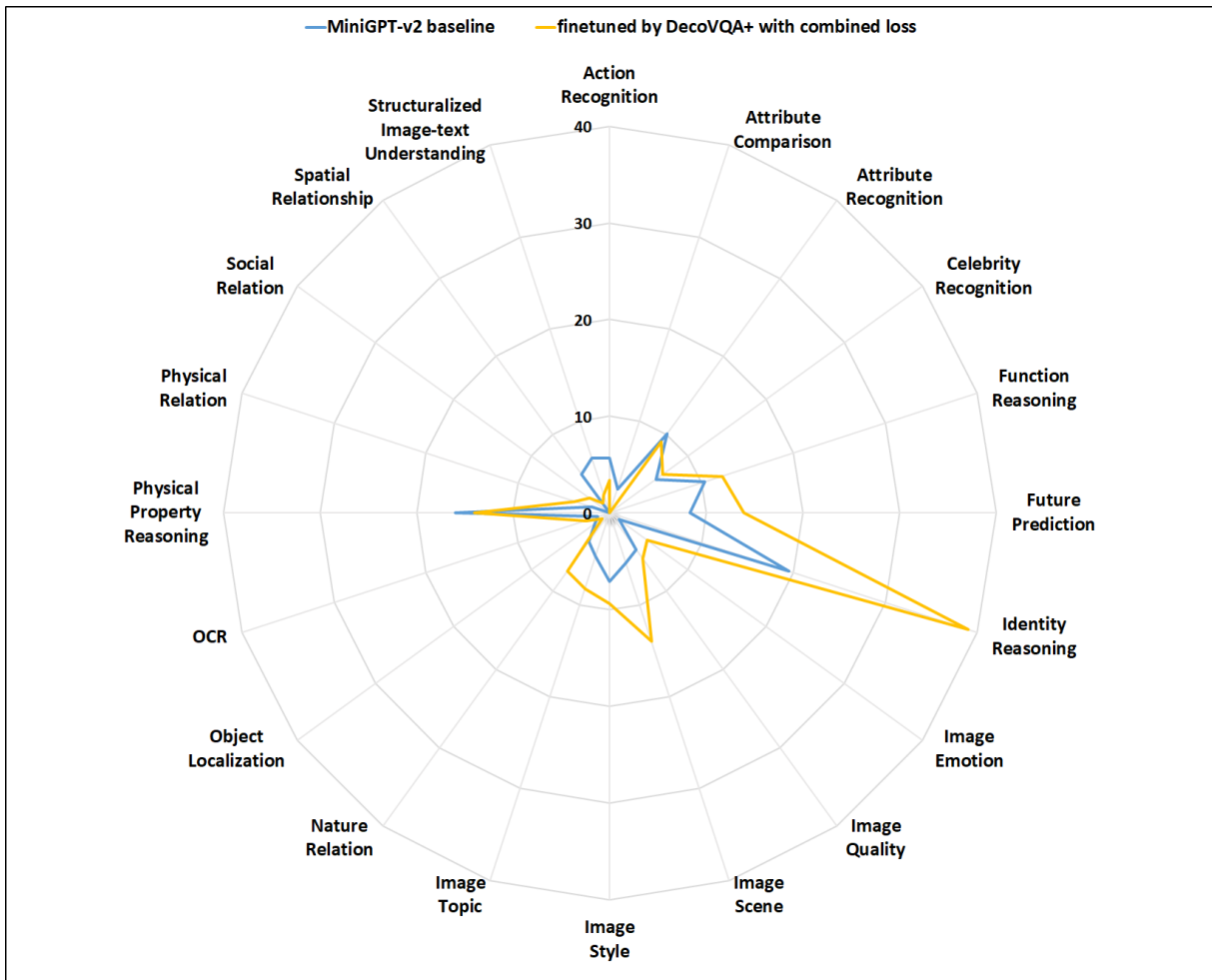


Figure 10: Results of different checkpoints of MiniGPT-v2 across the 20 L-3 ability dimensions defined in MMBench.

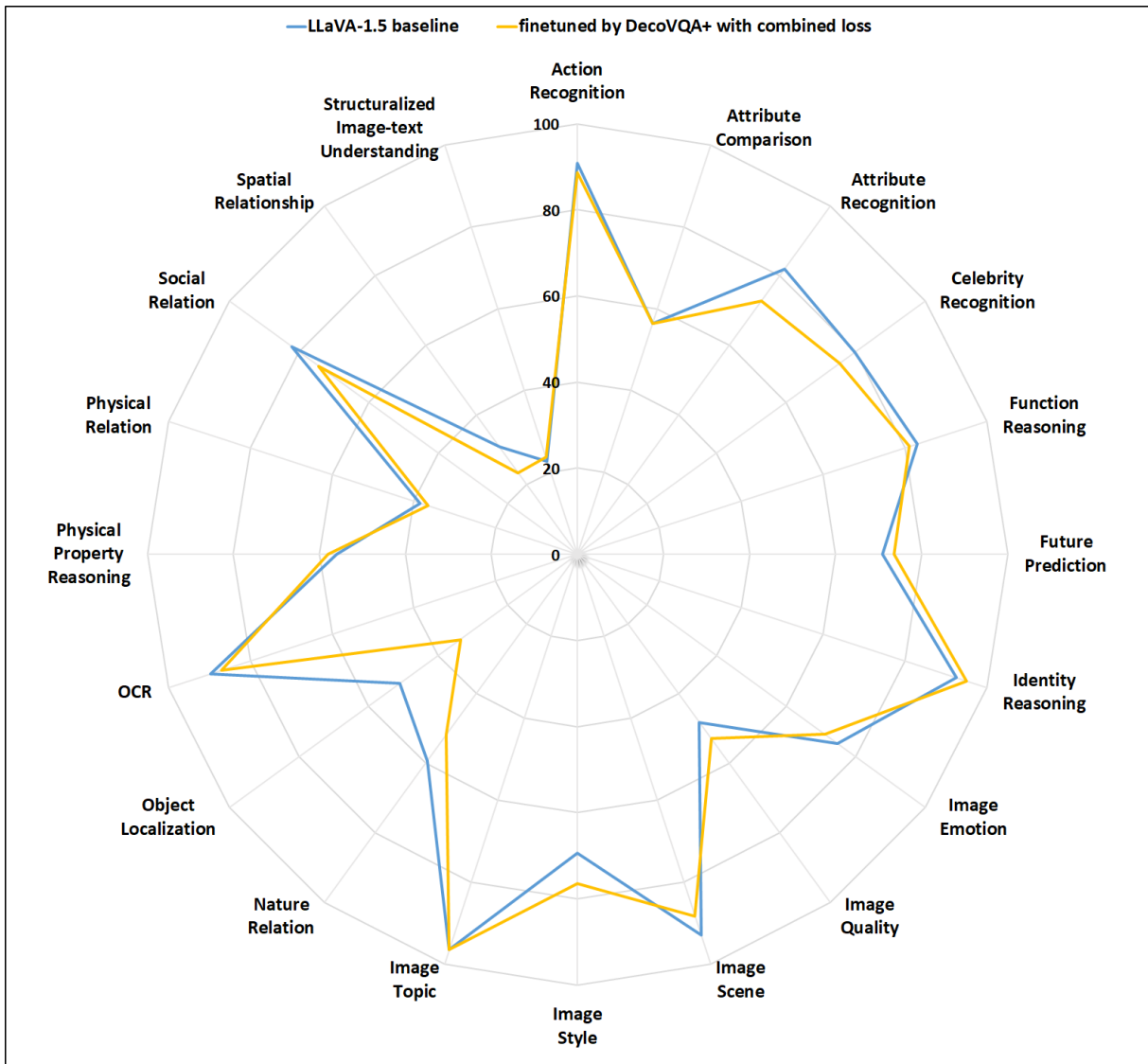


Figure 11: Results of different checkpoints of LLaVA-1.5 across the 20 L-3 ability dimensions defined in MMBench.

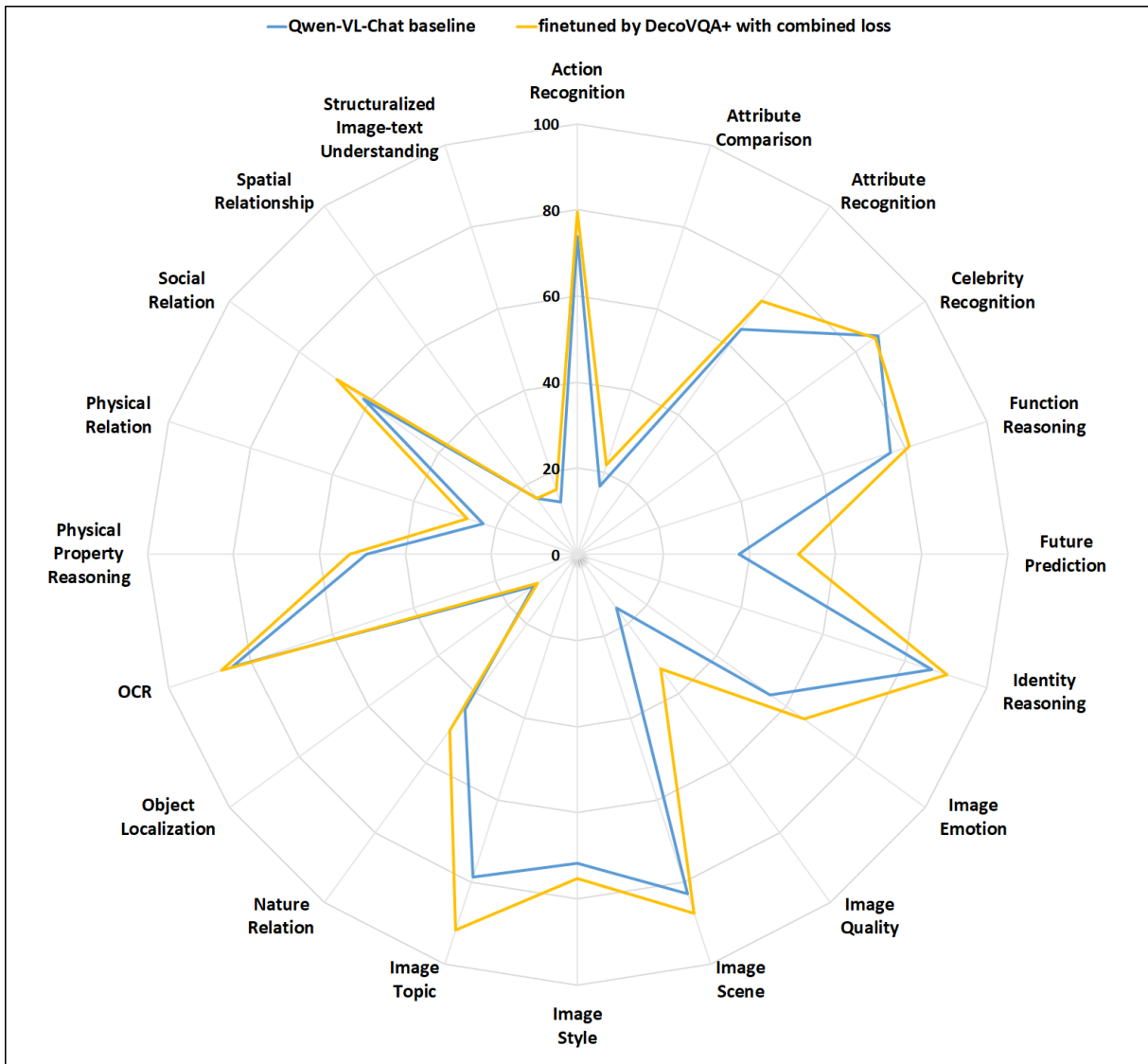


Figure 12: Results of different checkpoints of Qwen-VL-Chat across the 20 L-3 ability dimensions defined in MMBench.

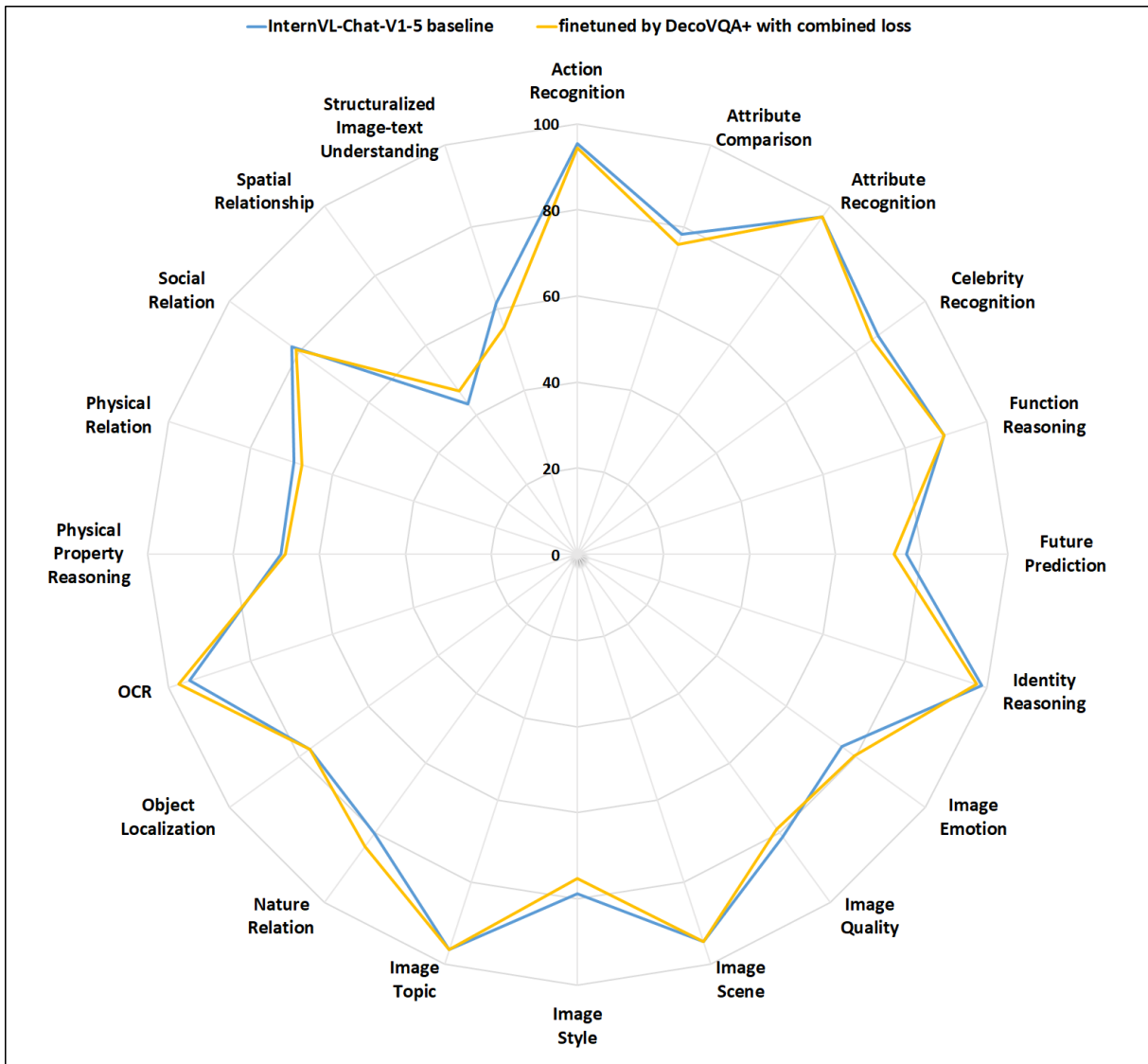


Figure 13: Results of different checkpoints of InternVL-Chat-V1-5 across the 20 L-3 ability dimensions defined in MMBench.

MiniGPT-v2	Finetuned by VQAIntrospect	Finetuned by DecoVQA+
Non-Repetition	17.20	88.35
Relevance	13.08	71.64
Groundedness	14.87	83.15
LLaVA-1.5	Finetuned by VQAIntrospect	Finetuned by DecoVQA+
Non-Repetition	21.52	94.18
Relevance	76.90*	78.67
Groundedness	93.50*	85.63

Table 10: Comparison of VQD abilities on MLLMs before and after finetuning with VQA-Introspect and with DecoVQA+. *Here for most of the original questions, LLaVA-1.5 produces one high quality sub-question, then repeats it for 2-3 times, causing relatively high score on Relevance and Groundedness, yet very low in Non-Repetition score.

Models	A-OKVQA	GQA	VQA-Introspect	Whether2Deco
MiniGPT-v2	41.2	44.2	62.1	46.8
finetuned by VQAIntrospect	48.8 ↑ (+7.6)	39.6 ↓ (-4.6)	63.7 ↑ (+1.6)	37.3 ↓ (-9.5)
finetuned by DecoVQA+	60.7 ↑ (+19.5)	50.7 ↑ (+6.5)	72.1 ↑ (+10.0)	61.0 ↑ (+14.2)
LLaVA-1.5	67.7	52.1	67.2	49.3
finetuned by VQAIntrospect	68.4 ↑ (+0.7)	51.8 ↓ (-0.3)	81.1 ↑ (+13.9)	4.8* ↓ (-44.5)
finetuned by DecoVQA+	72.7 ↑ (+5.0)	57.2 ↑ (+5.1)	75.4 ↑ (+8.2)	68.8 ↑ (+19.5)

Table 11: Comparison of VQA accuracy (%) on external knowledge (A-OKVQA) and visual reasoning (GQA and VQA-Introspect) datasets and Whether2Deco accuracy (%) before and after fine-tuning MLLMs with VQA-Introspect and with DecoVQA+. *Here LLaVA-1.5 fails to follow the pre-defined answering template, but to perform pure question decomposition instead of selective decomposition.

K Comparison with the Unimodal QD Method

Existing researches (You et al., 2023; Qi et al., 2023) tend to use a convincing captioning model to convert images to the language descriptions, and then perform the unimodal question decomposition with LLMs. Table 12 shows the accuracy gap under the selective VQD inference setting between MLLMs and their corresponding LLMs with GPT-4V as the captioning model. Since critical information in images is often lost during the captioning process, it is very possible for the subsequent inference with QD to fail to answer questions correctly. To sum up, VQD is better than the method "caption + QD".

L Comparison with In-context Learning Method

Besides finetuning, In-context Learning (ICL) is also a potential approach for VQD. The previous work (Khan et al., 2023) has explored VQD based on ICL methods. Therefore, we add an experiment to compare the performance with our finetuning pipeline and with the ICL method.

Models	VQA-Introspect
MiniGPT-v2	62.1
Llama2-Chat-7B-HF	46.2
LLaVA-1.5	67.2
Vicuna-13B-v1.5	62.3

Table 12: Comparison of VQA accuracy (%) between MLLMs and their corresponding language models on VQA-Introspect.

To fairly compare with the ICL method used in (Khan et al., 2023), we apply the same 2-shot demonstration as the one applied in that paper to decompose questions. The prompt template is shown in Figure 14.

The performance comparison in Table 13 and Table 14 shows that the models achieve significantly better performance in VQD ability, VQA accuracy and Whether-to-decompose accuracy through our finetuning pipeline, compared to the ICL method proposed in (Khan et al., 2023).

Model	Non-Repetition	Relevance	Groundedness
MiniGPT-v2 (zero-shot)	47.52	36.65	43.30
MiniGPT-v2 (ICL)	54.65	49.64	49.97
MiniGPT-v2 (finetuned by DecoVQA+)	88.35	71.64	83.15
LLaVA-1.5 (zero-shot)	42.19	37.33	44.17
LLaVA-1.5 (ICL)	69.45	65.32	62.58
LLaVA-1.5 (finetuned by DecoVQA+)	94.18	78.67	85.63

Table 13: Comparison of VQD ability between MLLMs finetuned by DecoVQA+ and inference with ICL-method across three evaluation criteria.

Model	A-OKVQA	GQA	VQA-Introspect	Whether2Deco
MiniGPT-v2 (zero-shot)	41.2	44.2	62.1	46.8
MiniGPT-v2 (ICL)	40.1	43.6	60.5	46.8
MiniGPT-v2 (finetuned by DecoVQA+)	64.0	51.7	72.5	71.5
LLaVA-1.5 (zero-shot)	67.7	52.1	67.2	49.3
LLaVA-1.5 (ICL)	65.1	51.3	67.4	49.3
LLaVA-1.5 (finetuned by DecoVQA+)	73.9	56.7	75.8	75.0

Table 14: Comparison of Accuracy (%) between MLLMs finetuned by DecoVQA+ and inference with ICL-method.

M More case studies

More case studies in addition to Figure 4 are shown in Figure 15.

N Licensing

Our proposed datasets SubQuestRater Dataset, DecoVQA, DecoVQA+, and Whether2Deco are built upon the public datasets A-OKVQA and VQA-Introspect. A-OKVQA has the Apache-2.0 License.

Prompt under ICL setting (two-shot)

Please firstly decompose the given question into several image-relevant sub-questions to help you answer the given question. Please avoid giving repeated sub-questions or generating an excessive number. Feel free to suggest an appropriate quantity based on your judgment. Here are two examples you can follow to decompose the question:

Example 1

Question: Is the banana ripe enough to eat?

Sub-questions: 1. Is the banana yellow?

Example 2

Question: Is it cold outside?

Sub-questions: 1. Are any people wearing jackets?




Input

Question: {question}




Sub-questions:

Figure 14: Prompt under ICL setting (two-shot)


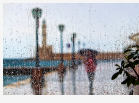

The licenses of the code for the mentioned MLLMs are listed as follows: MiniGPT-v2 has the BSD-3-Clause License, LLaVA-1.5 has the Apache-2.0 License, Qwen-VL-Chat has the Tongyi Qianwen License and InternVL-Chat-V1-5 has the MIT License. We publicize all of our proposed datasets and our code under the MIT License.

<p>User: Where can these foods be found? Choose from A) fast food, B) bar, C) garden, D) office</p> 	<p>User: What is required for this activity? Choose from A) water, B) snow, C) sun, D) wind</p> 	<p>User: Is this a safe airplane?</p> 
<p>original MiniGPT-v2</p> <p>A B C D</p>	<p>original MiniGPT-v2</p> <p>A, B, C</p>	<p>original MiniGPT-v2</p> <p>-Is the plane in flight? -What is its altitude? -How far from the airport? -What direction is it travelling?</p>
<p>MiniGPT-v2 finetuned by DecoVQA+</p> <p>1. What are the foods in the image? 2. Do these foods grow on trees or plants? 3. Are these foods commonly served at fast-food restaurants, bars, or offices?</p>	<p>MiniGPT-v2 finetuned by DecoVQA+</p> <p>1. What are the people in the image wearing? 2. Do the skis and poles the people are holding indicate what activity they are involved in? 3. Does the snow conditions indicate the people's activity?</p>	<p>MiniGPT-v2 finetuned by DecoVQA+</p> <p>1. What is the color and make of the airplane? 2. Is the airplane in the air? 3. Are all its engines running? 4. Is the plane's wings and body structurally sound?</p>




(a) Cases with MiniGPT-v2 before and after being finetuned by DecoVQA+.

<p>User: What type of store is this? Choose from A) shoe, B) grocery, C) pet, D) beauty</p> 	<p>User: What are the people waiting to do? Choose from A) eat, B) work, C) speak, D) cross</p> 	<p>User: What season can we infer it is?</p> 
<p>original LLaVA-1.5</p> <p>What type of store is this? Choose one option from A) shoe, B) grocery, C) pet, D) beauty</p> <p>1. What are the people in the store doing? 2. What are the people in the store looking at? 3. What are the people in the store selling?</p>	<p>original LLaVA-1.5</p> <p>1. What are the people waiting to do? A) eat B) work C) speak D) cross</p>	<p>original LLaVA-1.5</p> <p>1. What is the color of the teddy bear in the image? 2. What is the color of the teddy bear's scarf? 3. What is the color of the teddy bear's hat? 4. What is the color of the teddy bear's shirt? 5. What is the color of the teddy bear's pants? 6. What is the color of the teddy bear's shoes?</p>
<p>LLaVA-1.5 finetuned by DecoVQA+</p> <p>1. What are the people in the store doing? 2. What are the cages on the shelves? 3. Does the presence of cages indicate the type of store?</p>	<p>LLaVA-1.5 finetuned by DecoVQA+</p> <p>1. What are the people standing on? 2. What direction are the people facing? 3. What sign is on the post?</p>	<p>LLaVA-1.5 finetuned by DecoVQA+</p> <p>1. What is the color and make of the airplane? 2. Is the airplane in the air? 3. Are all its engines running? 4. Is the plane's wings and body structurally sound?</p>

(b) Cases with LLaVA-1.5 before and after being finetuned by DecoVQA+.

<p>User: How does the woman feel? Choose one option from A) cool, B) wet, C) cold, D) hot</p> 	<p>User: Why does the person have an umbrella? Choose one option from A) disguise, B) sun, C) snow, D) rain</p> 	<p>User: Does this kitchen appear to be under construction?</p> 
<p>original Qwen-VL-Chat</p> <p>A) How does the woman feel physically? B) Is the woman feeling damp or wet? C) Is the woman feeling cold or chilly? D) Is the woman feeling hot or warm?</p>	<p>original Qwen-VL-Chat</p> <p>A) Why does the person need to hide something? B) Why does the person need to protect themselves from the sun? C) Why does the person need to protect themselves from the snow? D) Why does the person need to protect themselves from the rain?</p>	<p>original Qwen-VL-Chat</p> <p>1. Is the image of a kitchen? 2. Does the kitchen appear to be under construction?</p>
<p>Qwen-VL-Chat finetuned by DecoVQA+</p> <p>1. What is the woman's facial expression? 2. Does the woman's facial expression indicate her emotional state? 3. Does the woman's facial expression indicate her emotional state?</p>	<p>Qwen-VL-Chat finetuned by DecoVQA+</p> <p>1. What is the weather? 2. Does the weather require an umbrella? 3. Does the person's clothing indicate the weather?</p>	<p>Qwen-VL-Chat finetuned by DecoVQA+</p> <p>1. What is the current state of the kitchen? 2. Is the kitchen being renovated? 3. Is the kitchen being built? 4. What is the condition of the kitchen?</p>

(c) Cases with Qwen-VL-Chat before and after being finetuned by DecoVQA+.

<p>User: What's the main property of the red material in the jar on the back table? Choose one option from A) sour, B) spicy, C) oily, D) salty</p> 	<p>User: Where are the people located? Choose one option from A) plateau, B) rainforest, C) savannah, D) town</p> 	<p>User: Is this a homemade meal?</p> 
<p>original InternVL-Chat-V1-5</p> <p>1. What is the red material in the jar? 2. What is the main property of the red material? 3. Is the red material sour? 4. Is the red material spicy? 5. Is the red material oily? 6. Is the red material salty?</p>	<p>original InternVL-Chat-V1-5</p> <p>1. What is the setting of the image? 2. Are there any identifiable landmarks or features in the image? 3. Can the location be determined based on the image?</p>	<p>original InternVL-Chat-V1-5</p> <p>1. What is in the picture? 2. Is the meal homemade?</p>
<p>InternVL-Chat-V1-5 finetuned by DecoVQA+</p> <p>1. What is the red material in the jar? 2. What is the main property of the red material?</p>	<p>InternVL-Chat-V1-5 finetuned by DecoVQA+</p> <p>1. Does the greenery and the building indicate that the people are located in a town? 2. Does the park setting indicate where the people are located?</p>	<p>InternVL-Chat-V1-5 finetuned by DecoVQA+</p> <p>1. What is the food on the plate? 2. Is the food on the plate a common breakfast? 3. Is the food on the plate a homemade meal?</p>

(d) Cases with InternVL-Chat-V1-5 before and after being finetuned by DecoVQA+.

Figure 15: Case studies showing the comparison of VQD performance by MLLMs before and after finetuning by DecoVQA+.