Decompose and Compare Consistency: Measuring VLMs' Answer Reliability via Task-Decomposition Consistency Comparison

Qian Yang[1,2], **Weixiang Yan**[3], **Aishwarya Agrawal**[1,2,4]
[1] Mila - Québec AI Institute [2] Université de Montréal
[3] University of California, Santa Barbara [4] Canada CIFAR AI Chair
qian.yang@mila.quebec weixiangyan@ucsb.edu aishwarya.agrawal@mila.quebec

Abstract

Despite tremendous advancements, current state-of-the-art Vision-Language Models (VLMs) are still far from perfect. They tend to hallucinate and may generate biased responses. In such circumstances, having a way to assess the reliability of a given response generated by a VLM is quite useful. Existing methods, such as estimating uncertainty using answer likelihoods or prompt-based confidence generation, often suffer from overconfidence. Other methods use self-consistency comparison but are affected by confirmation biases. To alleviate these, we propose **De**compose and **C**ompare **C**onsistency (DeCC) for reliability measurement. By comparing the consistency between the direct answer generated using the VLM's internal reasoning process, and the indirect answers obtained by decomposing the question into sub-questions and reasoning over the sub-answers produced by the VLM, DeCC measures the reliability of VLM's direct answer. Experiments across six vision-language tasks with three VLMs show DeCC's reliability estimation achieves better correlation with task accuracy compared to the existing methods. The code is publicly available at https://github.com/MyLittleChange/DeCC.

1 Introduction

Automatic measurement of reliability of responses generated by AI systems such as vision-language models (VLMs) is useful for deciding whether to trust a response or not, which in turn is necessary to build secure systems and enable further improvements (Varshney and Baral, 2023). Existing reliability estimation methods often estimate the model's uncertainty using answer likelihoods or prompt the model to generate a confidence value (Xiong et al., 2024; Tian et al., 2023; Mielke et al., 2022). These methods often fail to correlate well with task accuracy because models are not well-calibrated and tend to be overconfident (Chen et al., 2023b). Other methods attempt to incorporate calibrated confidence generation as a training goal (Lin et al., 2022; Ye and Durrett, 2022; Oh et al., 2024), but retraining the model is inefficient and even impractical for measuring the reliability of multiple VLMs or closed-source models. Some works use self-consistency to measure reliability by comparing the consistency among multiple generated answers (Wang et al., 2022; Chen et al., 2024a, 2023a), but self-consistency might suffer from confirmation biases (Feng et al., 2024).

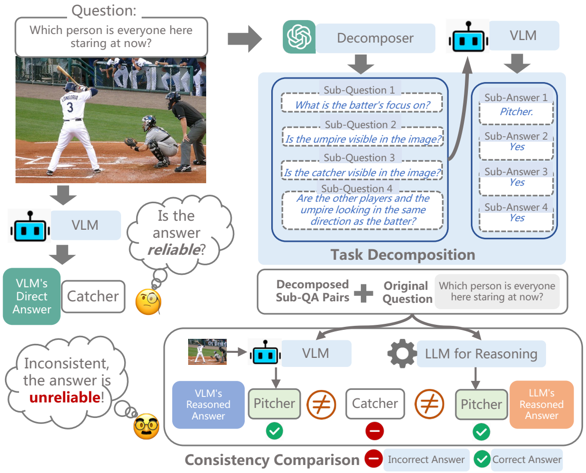


Figure 1: DeCC begins by decomposing the question into multiple sub-questions. The candidate VLM answers these sub-questions, creating sub-QA pairs. Both the candidate VLM and an LLM independently reason over these pairs to derive reasoned answers. We then compare the direct answer with the reasoned answers to assess reliability. We also explore how different consistency comparison settings impact DeCC's effectiveness.

To better measure VLMs' answer reliability, we propose a method called **De**compose and **C**ompare **C**onsistency (DeCC). As shown in Fig 1, we first decompose the original question into several sub-questions. The candidate VLM then answers these sub-questions, generating a sequence of sub-QA pairs. We use both the candidate VLM and a separate LLM, acting as two independent agents,

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 3613–3627
November 12-16, 2024 ©2024 Association for Computational Linguistics

to reason over the sub-QA pairs and obtain their respective reasoned answers. We then compare the consistency between these reasoned answers and the answer generated directly by the VLM to measure the reliability of the VLM's direct answer. Using the candidate VLM to reason over sub-QA pairs provides insights into how robustly the VLM understands the question. However, such self-consistency can sometimes introduce confirmation biases (Feng et al., 2024). Thus, we also employ an LLM to reason over the sub-QA pairs separately. We test both single-agent and multi-agent settings. For the single-agent setting, we use the consistency between the direct answer and one of the agent's reasoned answers to determine reliability. For the multi-agent setting, we combine the consistency check results from both agents to determine if the answer is reliable, unreliable, or requires further information for measurement. We assume that if the VLM understands the question well and conducts reliable reasoning, a conflict is less likely to occur between its direct answer, derived from its internal reasoning process, and the decomposed answer, derived from an external reasoning process. We evaluate DeCC on six vision-language tasks using three different state-of-the-art VLMs. Experimental results demonstrate that DeCC, which is both model-agnostic and task-agnostic, exhibits a higher correlation with the VLMs' task accuracy compared to the existing methods. Additionally, we observe that the effectiveness of different consistency comparison settings is correlated with the candidate VLM's capabilities.

2 Related Work

Existing methods use uncertainty-based metrics for reliability measurement, such as setting a reliability threshold on answer likelihoods (Pereyra et al., 2017; Geifman and El-Yaniv, 2017; Whitehead et al., 2022), or prompting the model to generate a confidence value (Xiong et al., 2024; Tian et al., 2023; Li et al., 2024; Mielke et al., 2022). However, uncertainty-based metrics often lead to overconfidence since confidence calibration is not a training goal (Chen et al., 2023b). But retraining models to generate calibrated confidence (Oh et al., 2024; Lin et al., 2022; Zhang et al., 2023) is impractical for evaluating multiple VLMs. Self-consistency methods generate multiple responses to assess reliability (Wang et al., 2022; Chen et al., 2024a, 2023a) but suffer from confirmation biases (Huang et al., 2024; Xie et al., 2024). Multi-agent collaboration can mitigate this. Feng et al. (2024) use multiple LLMs to interact in cooperative and competitive settings to evaluate reliability. Srinivasan et al. (2024) use LLMs to generate related questions about the image and use high-confidence QA pairs as premises, with the original QA as the hypothesis, to determine reliability. Our approach differs by decomposing the question into simpler sub-questions. We also conduct extensive experiments to explore the effectiveness of different consistency comparison settings on reliability measurement.

3 Method

For a question Q, an image I, and an answer A from a candidate VLM, DeCC obtains a binary Reliability score $\mathcal{R}$ indicating whether A is Reliable. As shown in Fig 1, DeCC contains two components: Task Decomposition and Consistency Comparison.

3.1 Task Decomposition

First, the decomposer, which could be any VLM, decomposes the question Q into a sequence of sub-questions conditioned on I. The candidate VLM then answers these sub-questions, resulting in a sequence of sub-QA pairs. Next, the candidate VLM and a separate LLM, acting as two independent agents, reason over the sub-QA pairs and Q, yielding VLM's reasoned answer $A_V^{'}$ and LLM's reasoned answer $A_L^{'}$. To enhance robustness, we also experiment with a two-iteration decomposition process. In the second iteration, sub-QA pairs from the first iteration, along with Q and I, are used to guide the decomposer in generating additional sub-questions. The candidate VLM answers these new sub-questions, conditioned on I and previous sub-QA pairs, resulting in new sub-QA pairs containing more information. Finally, both agents reason over all sub-QA pairs from both iterations to provide their updated reasoned answers, $A_V^{''}$ and $A_L^{''}$ [1].

3.2 Consistency Comparison

We explore both single-agent and multi-agent settings for consistency comparison to obtain $\mathcal{R}$.

Single-Agent We compare the VLM's direct answer A with either the VLM's reasoned answer $A_V^{'}$ (*VLM Agent Consistency*) or the LLM's reasoned

[1] Single quotation mark ($'$) annotates the first iteration and double quotation mark ($''$) annotates the second iteration.

answer $A_L^{'}$ (*LLM Agent Consistency*) and obtain:

$$\mathcal{R} = \begin{cases} 1, & \text{if } A^{'} \text{ is consistent with } A \\ 0, & \text{otherwise} \end{cases}$$

We check if $A^{'} = A$ to determine the consistency. For two-iteration decomposition, we compare A with $A_V^{''}$ and $A_L^{''}$ to obtain $\mathcal{R}$ in a similar way.

Multi-Agent As shown in Fig 2, we first conduct consistency checks of A with $A_V^{'}$ and $A_L^{'}$ and obtain $Cons_V^{'}$ (consistency between A and $A_V^{'}$) and $Cons_L^{'}$ (consistency between A and $A_L^{'}$). If $Cons_V^{'} = Cons_L^{'}$, we assign $\mathcal{R} = Cons_V^{'}$. If $Cons_V^{'} \neq Cons_L^{'}$, we proceed to the second-iteration consistency checks, where we compare updated reasoned answers $A_V^{''}$ and $A_L^{''}$ with A, obtaining $Cons_V^{''}$ and $Cons_L^{''}$. We assign $\mathcal{R}$ as:

$$\mathcal{R} = \begin{cases} Cons_V^{''}, & \text{if } Cons_V^{''} = Cons_L^{''} \\ Cons_L^{''}, & \text{if } Cons_V^{'} = Cons_V^{''} \text{ and} \\ & \quad Cons_L^{'} = Cons_L^{''} \\ Cons_V^{''}, & \text{if } Cons_V^{'} \neq Cons_V^{''} \text{ and} \\ & \quad Cons_L^{'} \neq Cons_L^{''} \end{cases}$$

(1) The first scenario indicates that the consistency check outcome for one of the agents has changed from the first iteration, leading to the same consistency check outcomes between the two agents. (2) The second scenario indicates that both agents show strong confidence in their respective consistencies with respect to the direct answer. We trust the LLM's consistency check, as it provides a more objective assessment, Relying solely on textual decomposition information, whereas the VLM might suffer from its inherent biases towards certain responses. (3) The third scenario indicates that the second-iteration decomposition provides additional information, influencing both agents' reasoning and changing their consistency with respect to the direct answer. We trust the VLM's consistency check outcome, as VLM is less likely to change its response due to its inherent biases, whereas the LLM's response is more likely to change since it is operating under incomplete information (lack of image). So a change in VLM's response indicates it potentially overcame its biases with additional sub-QA pairs. See Appendix for Algorithm 1.

4 Experiments

4.1 Datasets

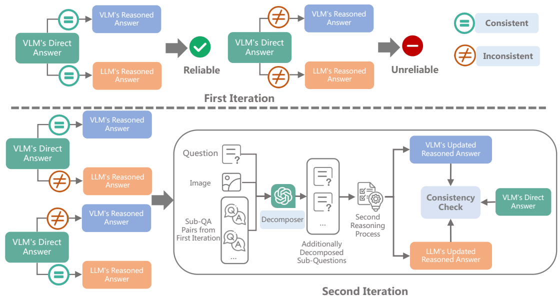


Figure 2: Illustration of Multi-Agent Consistency Comparison. *Top:* When both agents' reasoned answers are either consistent or inconsistent with the VLM's direct answer, we directly determine the Reliability. *Bottom:* If there is a contradiction in consistency check results, we proceed to the second-iteration consistency checks.

We conduct experiments on six VQA tasks: SNLI-VE (Xie et al., 2019), VCR (Zellers et al., 2019), A-OKVQA (Schwenk et al., 2022), Winoground (Thrush et al., 2022), MMMU (Yue et al., 2023), and MathVista (Lu et al., 2024). More details can be found on Appendix A.1.

4.2 Evaluation Metric

We use the Brier Score (BS)(Brier, 1950) to measure the correlation between reliability and task accuracy: BS $= \frac{1}{N}\sum_{i=1}^{N}(\mathcal{R}_i - Acc_i)^2$, where N is the evaluation dataset size, $\mathcal{R}_i$ is the reliability score for the i-th answer, and Acc_i is the accuracy for the i-th answer. BS ranges between 0 and 1, with lower values indicating better correlation between $\mathcal{R}$ and Acc. We also apply DeCC for the selective prediction task where the model abstains from answering when it's response is estimated to be unreliable. To measure DeCC effectiveness at selective prediction we use the Effective Reliability (ER) metric proposed in (Whitehead et al., 2022). ER captures the trade-off between risk (task accuracy across all answered questions) and coverage (number of questions answered). Both low risk but low coverage and high coverage but high risk lead to low ER. ER for the i-th answer is computed as:

$$ER(A_i) = \begin{cases} 1 & \text{if } \mathcal{R}_i = 1 \text{ and } Acc_i = 1 \\ -1 & \text{if } \mathcal{R}_i = 1 \text{ and } Acc_i = 0 \\ 0 & \text{if } \mathcal{R}_i = 0 \text{ (answer abstention)} \end{cases}$$

4.3 Existing Methods Used for Comparison

Perplexity of Direct Answer: Calculate the mean perplexity over tokens of the direct answer and use a threshold to determine reliability. If perplexity exceeds the threshold, $\mathcal{R}$ is 0 otherwise 1. **Generated Numerical Confidence**: Prompt the VLM to generate a confidence value along with the answer, formatted as 'Answer: X. Confidence:

Method	SNLI		VCR		A-OKVQA		Wino.		MMMU		MathVista		Mean	
	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑
LLaVA1.5-7B as Candidate VLM	*Acc: 55.0*		*Acc: 59.2*		*Acc: 67.3*		*Acc: 59.6*		*Acc: 34.3*		*Acc: 24.5*		*Acc: 49.9*	
Perplexity of Direct Answer	55.7	0.7	38.2	20.4	22.8	55.0	39.3	24.3	42.4	-8.2	**25.9**	**-1.4**	37.4	15.1
Generated Numerical Confidence	66.5	-32.5	40.8	18.3	22.1	55.6	28.0	44.0	67.3	-35.3	75.8	-51.6	50.1	-0.2
Generated Linguistic Confidence	67.5	-35.0	40.2	19.6	22.6	54.8	**27.6**	**44.8**	69.6	-39.1	77.2	-54.4	50.8	-1.5
Self-Consistency based on Paraphrase	38.5	17.5	32.8	25.7	19.0	59.2	40.5	23.9	39.1	-5.6	35.6	-11.5	34.3	18.2
DeCC														
VLM Agent Consistency	31.9	24.5	36.4	22.2	**18.2**	**59.6**	35.3	28.3	52.3	-18.1	46.3	-21.8	36.7	15.8
VLM Agent Consistency (2 iterations)	32.5	23.9	34.5	24.1	18.3	59.5	36.1	27.4	49.1	-14.9	45.6	-21.1	36.0	16.5
LLM Agent Consistency	32.0	24.4	35.9	22.7	24.5	53.3	37.5	26.0	**34.1**	**0.1**	30.7	-6.2	32.4	20.1
LLM Agent Consistency (2 iterations)	**30.6**	**25.8**	**32.6**	**26.0**	22.3	55.5	34.6	28.9	36.8	-2.6	31.0	-6.5	**31.3**	**21.2**
Multi-Agent Consistency (2 iterations)	31.5	24.9	33.5	25.1	20.1	57.7	34.6	28.9	36.4	-2.2	32.2	-7.7	31.4	21.1
Idefics2-8B as Candidate VLM	*Acc: 39.3*		*Acc: 78.6*		*Acc: 83.1*		*Acc: 70.0*		*Acc: 39.9*		*Acc: 48.0*		*Acc: 59.8*	
Perplexity of Direct Answer	59.7	-20.0	34.1	28.2	19.9	63.2	29.8	43.5	40.6	-1.0	30.0	15.1	35.6	21.5
Generated Numerical Confidence	40.8	-0.5	37.7	25.3	36.3	46.7	25.3	49.1	67.7	-43.6	49.3	-1.6	42.8	12.6
Generated Linguistic Confidence	35.0	-3.1	40.2	22.1	25.2	56.6	26.8	45.6	60.4	-36.3	42.4	3.5	38.3	14.7
Self-Consistency based on Paraphrase	59.1	-19.3	31.6	30.4	16.3	66.5	28.9	43.8	41.6	-2.0	40.8	4.8	36.4	20.7
DeCC														
VLM Agent Consistency	44.9	-5.2	30.5	31.6	13.9	69.2	22.6	50.4	43.9	-4.4	**28.7**	15.5	30.8	26.2
VLM Agent Consistency (2 iterations)	47.8	-8.1	**29.5**	**33.1**	**13.8**	**69.3**	**22.3**	**50.9**	43.0	-3.6	29.4	**15.9**	31.0	26.3
LLM Agent Consistency	**34.3**	5.5	37.9	24.4	26.3	56.5	35.3	38.0	**34.2**	**5.3**	40.8	4.4	34.8	22.3
LLM Agent Consistency (2 iterations)	34.9	**6.3**	34.0	25.0	24.0	61.4	32.0	39.3	35.9	5.1	34.0	11.4	32.5	24.8
Multi-Agent Consistency	34.7	5.8	33.0	27.9	19.6	65.5	29.5	44.1	35.1	5.0	31.1	13.5	**30.5**	**27.0**
InternVL1.5-25.5B as Candidate VLM	*Acc: 70.2*		*Acc: 70.5*		*Acc: 88.5*		*Acc: 78.6*		*Acc: 43.7*		*Acc: 56.0*		*Acc: 67.9*	
Perplexity of Direct Answer	**28.0**	**42.2**	**27.5**	**43.6**	12.1	76.4	24.0	56.1	37.3	6.3	36.5	18.7	27.6	40.6
Generated Numerical Confidence	37.8	30.2	42.2	21.2	21.2	62.0	19.0	**62.1**	64.6	-29.4	39.6	17.6	37.4	27.3
Generated Linguistic Confidence	58.4	-26.0	31.4	37.9	15.7	68.6	43.4	13.3	71.6	-43.3	43.1	10.4	43.9	10.2
Self-Consistency based on Paraphrase	30.1	40.1	28.1	43.0	**11.0**	**77.5**	21.1	59.0	48.8	-5.0	52.9	3.6	32.0	36.4
DeCC														
VLM Agent Consistency	33.2	37.0	28.3	42.8	11.9	76.6	18.9	61.3	44.9	-1.2	**23.8**	**31.4**	**26.8**	**41.3**
VLM Agent Consistency (2 iterations)	33.9	36.3	29.1	42.0	11.3	77.2	**18.6**	61.5	44.8	-1.1	24.3	30.9	27.0	41.1
LLM Agent Consistency	36.3	33.9	37.6	33.5	22.2	66.3	29.4	50.8	40.3	3.3	37.1	18.1	33.8	34.3
LLM Agent Consistency (2 iterations)	34.5	35.7	34.9	36.2	18.8	69.7	27.0	53.1	**36.9**	**6.8**	33.3	21.9	30.9	37.2
Multi-Agent Consistency (2 iterations)	34.3	35.9	32.6	38.5	15.4	73.1	23.8	56.4	37.4	6.2	31.1	24.1	29.1	39.0

Table 1: Measuring Brier Score (**BS**) and Effective Reliability (**ER**) for various reliability measurement methods. Best results are in **bold**. Second-best results are underlined. *Acc* represents the task accuracy of the candidate VLM. All scores are in percentage. DeCC surpasses all baselines in average Brier Score and Effective Reliability.

X%'. A threshold determines reliability. **Generated Linguistic Confidence**: Prompt the VLM to state 'I am confident/not confident in this answer.' **Self-Consistency based on Paraphrase**: Prompt a VLM to paraphrase the original question into four variations. If n or more paraphrased answers differ from the direct answer, $\mathcal{R}$ is 0 otherwise 1. [2]

4.4 Main Results

We conduct experiments on six vision-language tasks[3], covering commonsense reasoning, fine-grained compositional reasoning, and science understanding (see Appendix A.1 for dataset descriptions). We evaluate three state-of-the-art VLMs: LLaVA1.5-7B (Liu et al., 2023), Idefics2-8B (Laurençon et al., 2024), and InternVL1.5-25.5B (Chen et al., 2024b) (see Appendix A.2 for implementation details). The overall results are shown in Table 1. DeCC achieves the best and second-best mean performance (mean across datasets) on Brier Score and Effective Reliability. DeCC reduces the relative mean Brier Score by 8.7% on LLaVA, 14.3% on Idefics2, and 2.9% on InternVL compared to the best existing methods. DeCC also increases relative mean Effective Reliability by 16.5% on LLaVA, 25.6% on Idefics2, and 1.7% on InternVL. We observe that with increasing VLM size, the performance of most methods improves, suggesting that reliability measurement is correlated with VLMs' capabilities. For the effectiveness of DeCC's different consistency comparison settings, we observe an interesting trend: (1) For weaker VLMs, i.e., LLaVA, LLM Agent Consistency achieves the best performance, likely because VLMs struggle to reason over the sub-QA pairs and suffer from confirmation biases. (2) For stronger VLMs, i.e. Idefics2, Multi-Agent Consistency performs the best suggesting that the VLM and LLM reasoners comple-

[2] We select the best threshold and n for each VLM based on the Brier Score (results in Tables 4 and 5).

[3] All datasets are multiple-choice except for MMMU and MathVista, whose answers are very short. We use string matching for consistency comparison.

Method	VCR		AOKVQA		Winoground		Mean	
	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑	BS↓	ER↑
Perplexity of Direct Answer	34.1	28.2	19.9	63.2	29.8	43.5	27.9	45.0
Generated Numerical Confidence	37.7	25.3	36.3	46.7	25.3	49.1	33.1	40.4
Generated Linguistic Confidence	40.2	22.1	25.2	56.6	26.8	45.6	30.7	41.4
Self-Consistency based on Paraphrase	31.6	30.4	16.3	66.5	28.9	43.8	25.6	46.9
DeCC - Greedy Decoding								
VLM Agent Consistency	**30.5**	**31.6**	13.9	69.2	**22.6**	**50.4**	**22.3**	50.4
LLM Agent Consistency	37.9	24.4	26.3	56.5	35.3	38.0	33.2	39.6
MultiAgent Consistency	33.0	27.9	19.6	65.5	29.5	44.1	27.4	45.8
DeCC - Sampling Decoding (Temp = 0.8)								
VLM Agent Consistency	31.2	31.1	**13.2**	**69.9**	22.9	50.3	22.4	**50.5**
LLM Agent Consistency	38.1	24.2	24.6	58.5	35.1	38.2	32.6	40.3
MultiAgent Consistency	32.7	29.6	18.5	64.6	27.8	47.2	26.3	47.1
DeCC - Sampling Decoding (Temp = 0.9)								
VLM Agent Consistency	31.7	30.6	14.1	69.0	23.3	49.9	23.0	49.9
LLM Agent Consistency	38.6	23.7	26.8	56.3	33.8	39.5	33.0	39.9
MultiAgent Consistency	33.7	28.6	20.9	63.3	28.4	44.9	27.7	45.6

Table 2: Results of difference decoding strategy using Idefics2-8B as the candidate VLM. Best results are in **bold**. Second-best results are underlined. In all decoding methods and temperatures, DeCC outperforms the baselines, demonstrating its effectiveness.

Benchmark	Questions per Sample	Question Types per Sample
SNLI	4.1	2.1
VCR	5.2	2.2
AOKVQA	4.4	2.1
Winoground	4.3	2.5
MMMU	5.9	2.4
MathVista	4.4	2.3

Table 3: Overview of the number of questions and question types per sample across different benchmarks.

ment each other. (3) For the strongest VLMs, i.e. InternVL, VLM Agent Consistency (self-consistency) achieves the best performance, as the VLM can effectively leverage the information contained in sub-QA pairs. Overall, the effectiveness of different consistency comparison settings correlates with the candidate VLM's capabilities.

4.5 Further Analysis

Decoding Strategy The decoding strategy of decomposer can influence the generated sub-questions, which in turn impacts DeCC's overall performance. We conducted experiments using sampling-based decoding with two different temperature settings, 0.8 and 0.9, while keeping the nucleus sampling (Holtzman et al., 2020) probability fixed at 0.9 for both. The results are presented in Table 2, using Idefics2-8B as the candidate VLM. Across all decoding strategies and temperature settings, DeCC consistently outperforms the baselines, demonstrating its effectiveness.

Question Type Analysis Previous studies have shown that VLMs can exhibit biases toward certain question types. For instance, they respond "yes" to over 80% of queries about non-existent objects, such as in prompts like "Is there an object in this image?" (Wang et al., 2023). If the decomposed sub-questions retain the same type as the original question, they may inherit the same bias, affecting evaluation reliability. To investigate this, we conduct the question type analysis to assess whether DeCC is prone to this issue [4]. Table 3 shows the number of questions and question types per sample across benchmarks. DeCC generates at least 2.1 distinct question types per sample for all benchmarks, reducing the impact of VLM biases.

Additional Analysis To better understand the workflow of DeCC, we provide some qualitative examples in Appendix A.4. We also provide the computational costs analysis in Appendix A.5 to address concerns about the additional computational requirements of DeCC and demonstrate its practical applicability.

5 Conclusion

We use consistency comparison based on task decomposition for measuring VLMs answer reliability. By decomposing complex questions into simpler sub-questions, we achieve more accurate and robust reliability estimation. We find the performance of reliability measurement and the effectiveness of different consistency comparison settings correlate with candidate VLM's capabilities.

[4] We categorize questions into types using string matching, with categories including: yes/no, color, number, how, why, what/which, when, where, who, and others.

Acknowledgement

We are grateful to Mila's IDT team for their technical support with the computational infrastructure. We thank Rabiul Awal for his constructive feedback. During this project, Aishwarya Agrawal was supported by the Canada CIFAR AI Chair award.

Limitations

Our experiments demonstrate that consistency comparison based on task decomposition can better measure the reliability of VLM answers. However, there are several limitations to our current study: *Decomposition Performance*: The effectiveness of our framework is influenced by the performance of the decomposition process. Currently, we have not fully explored the optimization and impact of different decomposition strategies for reliability measurement. *Multi-Agent Consistency Comparison*: We tested decomposition with only one LLM for the multi-agent part. Conducting more experiments with various LLMs will help assess the generalization and robustness of our framework. Future work will address these limitations to validate and enhance the generalization of our proposed method.

References

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. 2024a. Two failures of self-consistency in the multi-step reasoning of LLMs. *Transactions on Machine Learning Research*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2023a. Inside: Llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023b. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *Preprint*, arXiv:2402.00367.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Xiang Lisa Li, Urvashi Khandelwal, and Kelvin Guu. 2024. Few-shot recalibration of language models. *arXiv preprint arXiv:2403.18286*.

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

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

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

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

Changdae Oh, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. 2024. Towards calibrated robust fine-tuning of vision-language models. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions.

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

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Raghavi Chandu. 2024. Selective" selective prediction": Reducing unnecessary abstention in vision-language reasoning. *arXiv preprint arXiv:2402.15610*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Neeraj Varshney and Chitta Baral. 2023. Post-abstention: Towards reliably re-attempting the abstained instances in qa. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 967–982.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, Dublin, Ireland. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2023. R-tuning: Teaching large language models to refuse unknown questions. *arXiv preprint arXiv:2311.09677*.

A Experiments

A.1 Datasets

SNLI-VE requires VLMs to identify whether the relationship between the given image premise and text hypothesis is entailment, neutral, or contradiction. **Visual Commonsense Reasoning (VCR)** requires higher-order cognition and commonsense reasoning of VLMs. It provides an image and a question about certain objects in the image, along with four candidate answers, where the VLMs need to choose the correct answer. We add rectangles of different colors to the image and indicate the corresponding object's index in the upper right corner of each rectangle to distinguish the objects. **A-OKVQA** is an augmented successor of OK-VQA (Marino et al., 2019) and requires a broad base of commonsense and world knowledge to answer questions. Four candidate answers are provided along with each question. **Winoground (Wino.)** is proposed for measuring vision-linguistic compositional reasoning. It contains two images and two captions. The model needs to correctly match the captions to the images, but crucially, both captions contain an identical set of words, only in a different order. **MMMU** is designed to evaluate VLMs on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. Several candidate answers are provided along with each question. **MathVista** focuses on mathematical reasoning in visual contexts. We treat all datasets except for MMMU and MathVista as multiple-choice QA tasks. For evaluation:

- For SNLI-VE, VCR, and A-OKVQA, we randomly select 1,000 samples from the validation set.
- For Winoground, we feed one image and two captions to the VLM, which must correctly identify the corresponding caption, using a total of 800 samples.
- For MMMU, we evaluate on the validation set, which contains 900 samples.
- For MathVista, we evaluate on the testmini set, which contains 1,000 samples.

A.2 Implement Details

We use InternVL-1.5 (Chen et al., 2024b) as the decomposer for decomposition and question paraphrasing. For decomposition, we employ few-shot prompting by randomly selecting four samples from SNLI-VE and ScienceQA, with manually written decomposition processes as guidance. The few-shot prompt for decomposition is provided in Table 7. Only text is used in the few-shot prompt, without images. The decomposer determines the number of sub-questions needed. The few-shot prompt for the second-iteration decomposition is shown in Table 8 For paraphrasing, we use the same samples with manually written paraphrased questions. The few-shot prompt for paraphrasing is provided in Table 9. The remaining datasets are approached with a zero-shot strategy. We use OpenHermes-2.5-Mistral-7B[5] as the LLM Agent for reasoning. We evaluate three VLMs: LLaVA1.5-7B (Liu et al., 2023), Idefics2-8B (Laurençon et al., 2024), and InternVL (Chen et al., 2024b), all operating under a zero-shot setting across all datasets. Since all datasets are multiple-choice QA tasks or short answers, we use string matching for answer consistency. For baseline threshold settings:

- *Perplexity of Direct Answer*: 1.10 for LLaVA1.5-7B, 1.25 for Idefics2-8B, and 1.40 for InternVL based on Table 4.
- *Generated Numerical Confidence*: We set the threshold to 80%. If the generated confidence score exceeds 80%, the reliability score is 1; otherwise, it is 0.
- *Self-Consistency based on Paraphrase*: The number of inconsistent paraphrased-direct answer pairs is set to 0 for LLaVA1.5-7B and Idefics2-8B, and 2 for InternVL based on Table 5.

A.3 Evaluation Metric Selection

In our settings, we obtain binary reliability scores for each answer. We use the Brier Score (Brier, 1950) and Effective Reliability (Whitehead et al., 2022) to evaluate the reliability measurement. We do not use Expected Calibration Error (ECE) (Guo et al., 2017) because ECE is suitable for evaluating scores over a range of values. ECE relies on having a range of predicted probabilities to compare against actual accuracy. With only two reliability levels (0 or 1), there are no intermediate

[5] `https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B`

Metric	SNLI	VCR	A - OKVQA	Wino.	MMMU	MathVista	Mean
LLaVA							
Perplexity Threshold - 1.0	56.4	58.6	77.8	63.5	34.2	24.5	52.5
Perplexity Threshold - 1.05	56.4	47.4	36.0	58.1	31.5	24.6	42.3
Perplexity Threshold - 1.10	56.4	43.3	28.7	48.4	32.1	24.7	38.9
Perplexity Threshold - 1.15	56.2	41.9	25.1	41.3	35.2	25.1	37.5
Perplexity Threshold - 1.20	56.3	39.7	23.4	41.0	39.5	25.3	37.5
Perplexity Threshold - 1.25	55.7	38.2	22.8	39.3	42.4	25.9	**37.4**
Idefics2							
Perplexity Threshold - 1.0	39.7	62.3	83.1	73.3	40.0	45.1	57.2
Perplexity Threshold - 1.05	59.1	33.3	22.6	32.6	36.6	31.6	35.9
Perplexity Threshold - 1.10	59.7	34.1	19.9	29.8	40.6	30.0	**35.6**
Perplexity Threshold - 1.15	60.1	36.5	18.5	27.9	43.9	31.0	36.3
Perplexity Threshold - 1.20	60.3	37.5	17.0	27.0	49.0	32.4	37.2
Perplexity Threshold - 1.25	60.2	38.0	16.6	26.6	53.0	35.0	38.2
InternVL							
Perplexity Threshold - 1.0	70.2	71.1	88.5	80.2	43.6	55.2	68.1
Perplexity Threshold - 1.05	44.9	44.6	23.1	44.6	41.4	44.5	40.5
Perplexity Threshold - 1.10	38.8	38.0	17.9	37.1	39.2	40.8	35.3
Perplexity Threshold - 1.15	34.3	34.9	15.6	34.3	38.6	38.7	32.7
Perplexity Threshold - 1.20	31.8	32.5	14.1	31.3	38.9	35.4	30.7
Perplexity Threshold - 1.25	29.6	30.2	13.5	29.4	37.7	36.3	29.4
Perplexity Threshold - 1.30	28.3	29.1	12.7	27.5	36.6	36.1	28.4
Perplexity Threshold - 1.35	27.8	28.3	12.9	26.8	36.4	36.2	28.1
Perplexity Threshold - 1.40	28.0	27.5	12.1	24.0	37.3	36.5	**27.6**

Table 4: Brier Score using different threshold of perplexity on different VLMs. Best results are in **bold**. All scores are in percentage.

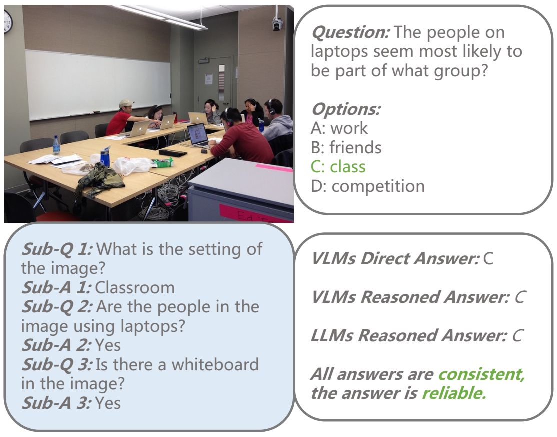
Metric	SNLI	VCR	A- OKVQA	Wino.	MMMU	MathVista	Mean
LLaVA							
Paraphrased Inconsistent - 0	38.5	32.8	19.0	40.5	39.1	35.6	**34.3**
Paraphrased Inconsistent - 1	39.5	34.1	19.2	37.1	46.6	44.1	36.8
Paraphrased Inconsistent - 2	41.2	36.4	19.9	37.6	50.0	49.7	39.1
Idefics2							
Paraphrased Inconsistent - 0	59.1	31.6	16.3	28.9	41.6	40.8	**36.4**
Paraphrased Inconsistent - 1	60.4	31.5	15.8	28.0	46.4	41.4	37.3
Paraphrased Inconsistent - 2	61.1	31.6	16.1	27.8	47.4	43.9	38.0
InternVL							
Paraphrased Inconsistent - 0	31.4	29.1	12.7	23.8	44.8	55.5	32.9
Paraphrased Inconsistent - 1	30.3	28.4	10.8	21.4	47.9	54.0	32.1
Paraphrased Inconsistent - 2	30.1	28.1	11.0	21.1	48.8	52.9	**32.0**

Table 5: Brier Score using different numbers of inconsistent paraphrased-direct answer pairs out of a total of 4 pairs. Best results are in **bold**. All scores are in percentage.

probabilities to assess the correlation. We also find Coverage at Risk (C@R) (Whitehead et al., 2022) not applicable to our settings. C@R measures the **C**overage proportion of correctly answered questions if we tolerate an **R**% of wrong answers by sorting predictions in descending order of score list and calculating coverage until the risk threshold is reached. C@R is not suitable for binary reliability scores because it relies on a range of reliability levels to sort and progressively evaluate predictions. With only binary scores, there is no meaningful way to sort the predictions by reliability. Consequently, C@R cannot provide a useful measure of performance in our setting.

A.4 Case Study

Fig 3 shows an example from A-OKVQA where all answers are consistent, and we assign the direct answer as reliable. Fig 4 shows an example from A-OKVQA where there is a contradiction between the consistency check results of the agents' reasoned answers and the direct answer. In this case, for the first sub-QA pair, the candidate VLM correctly identifies the birds as geese but fails to conduct correct reasoning over the decomposition process, deriving the same answer as the direct answer. Meanwhile, the LLM effectively utilizes the information from the decomposition. Both agents do not change their consistency check results. As illustrated in Section 3.2, we trust the LLM's con-



Question: The people on laptops seem most likely to be part of what group?

Options:
A: work
B: friends
C: class
D: competition

Sub-Q 1: What is the setting of the image?
Sub-A 1: Classroom
Sub-Q 2: Are the people in the image using laptops?
Sub-A 2: Yes
Sub-Q 3: Is there a whiteboard in the image?
Sub-A 3: Yes

VLMs Direct Answer: C

VLMs Reasoned Answer: *C*

LLMs Reasoned Answer: *C*

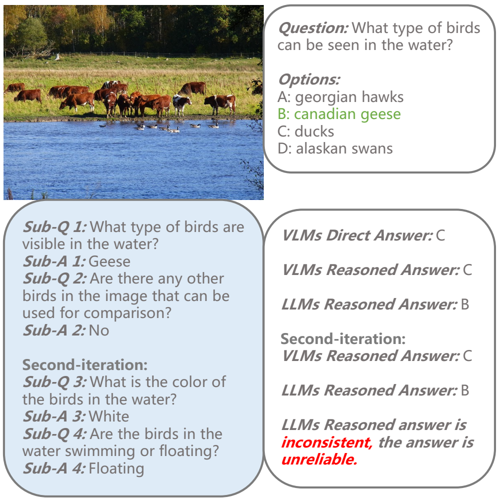
All answers are consistent, the answer is reliable.

Figure 3: Example for the consistent situation. All answers are consistent, thus we assign the direct answer as reliable.

sistency check results and assign the direct answer as unreliable. Fig 5 shows an example from VCR where all answers are inconsistent and incorrect, indicating that the VLMs do not understand the question well. We assign the direct answer as unreliable.

A.5 Computational Costs Analysis

DeCC requires multiple steps – *question decomposition, answering the decomposed questions, comparing consistency between the answers*. This raises concerns about DeCC being computationally more expensive than other approaches such as perplexity, generated numerical confidence scores, or self-consistency. To address this concern, we provide detailed computational costs for DeCC and compare it with that of baselines. Since the computational cost in DeCC is influenced by the number of sub-questions generated, we conduct this analysis on the VCR and AOKVQA, as these two datasets represent the median number of generated sub-questions among all benchmarks. All statistics are computed using Idefics2-8B (Laurençon et al., 2024) as the candidate VLM over a randomly chosen sample of 1,000 instances with a mean of three runs. For LLM Agent reasoning, we use vLLM (Kwon et al., 2023) for speeding up inference speed. Others are implemented using the Transformers toolkit[6] from Hugging Face without any speeding-up strategies. The consistency comparison stage requires only string comparisons, which is very fast (1 to 2 seconds for 1000 samples), so we do not list the time. We report the computational costs analysis in Table 6. For baselines such as perplexity of direct answer and generated confidence, the time cost is equal to direct answering, i.e., 0.22 seconds per sample for VCR and 0.16 seconds per sample for AOKVQA. For the baseline Self-Consistency based on Paraphrase, the consumed time is similar to the first iteration of DeCC, 4.8 seconds per sample for VCR and 3.9 seconds per sample for AOKVQA. For the first iteration of DeCC, the most time-consuming stage is the Question Pre-Decomposition stage for generating the sub-questions, taking 3.96 seconds per sample for VCR and 3.36 seconds per sample for AOKVQA. However, this process needs to be performed only once per sample before evalu-



Question: What type of birds can be seen in the water?

Options:
A: georgian hawks
B: canadian geese
C: ducks
D: alaskan swans

Sub-Q 1: What type of birds are visible in the water?
Sub-A 1: Geese
Sub-Q 2: Are there any other birds in the image that can be used for comparison?
Sub-A 2: No

Second-iteration:
Sub-Q 3: What is the color of the birds in the water?
Sub-A 3: White
Sub-Q 4: Are the birds in the water swimming or floating?
Sub-A 4: Floating

VLMs Direct Answer: C

VLMs Reasoned Answer: C

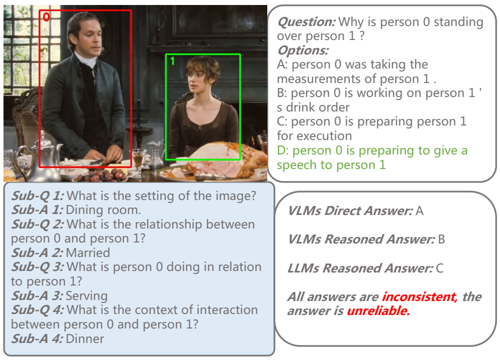
LLMs Reasoned Answer: B

Second-iteration:
VLMs Reasoned Answer: C

LLMs Reasoned Answer: B

LLMs Reasoned answer is inconsistent, the answer is unreliable.

Figure 4: Example for the inconsistent situation. The VLM's reasoned answer is consistent with the direct answer, while the LLM's reasoned answer is inconsistent. Both agents do not change their consistency check results. We trust the LLM's consistency check results and assign the direct answer as unreliable.



Question: Why is person 0 standing over person 1 ?
Options:
A: person 0 was taking the measurements of person 1 .
B: person 0 is working on person 1 ' s drink order
C: person 0 is preparing person 1 for execution
D: person 0 is preparing to give a speech to person 1

Sub-Q 1: What is the setting of the image?
Sub-A 1: Dining room.
Sub-Q 2: What is the relationship between person 0 and person 1?
Sub-A 2: Married
Sub-Q 3: What is person 0 doing in relation to person 1?
Sub-A 3: Serving
Sub-Q 4: What is the context of interaction between person 0 and person 1?
Sub-A 4: Dinner

VLMs Direct Answer: A

VLMs Reasoned Answer: B

LLMs Reasoned Answer: C

All answers are inconsistent, the answer is unreliable.

Figure 5: Example for the inconsistent situation. All answers are inconsistent, while none of these answers are correct, indicating the VLMs do not understand the question well. We assign the direct answer as unreliable.

[6] https://huggingface.co/docs/transformers/en/index

Method	**Sub-Questions/**	**Total Samples**	**Time per Sample**
	Paraphrased-Questions per Sample	**Used**	**(Seconds)**
VCR			
Baselines			
Perplexity of Direct Answer, Generated Confidence	N/a	1000	0.22
Self-Consistency based on Paraphrase	4	1000	4.80
DeCC Multi-Agent Consistency Stages			
Question Pre-Decomposition	3.9	1000	3.96
Sub-Question Answering and VLM Agent Reasoning	3.9	1000	0.84
LLM Agent Reasoning	3.9	1000	0.18
Second Question Decomposition	4.1	366	4.09
Second Sub-Question Answering and VLM Agent Reasoning	4.1	366	0.93
Second LLM Agent Reasoning	4.1	366	0.20
DeCC Expected Value (Multi-Agent Consistency)	$[1000 \times 3.9 + 366 \times 4.1]/1000 = 5.4$	1000	$[1000 \times (3.96 + 0.84 + 0.18) +$ $366 \times (4.09 + 0.93 + 0.20)]/1000 = 6.89$
DeCC Expected Value (VLM-Agent Consistency (2 Iterations))	8.0	1000	9.82
DeCC Expected Value (LLM-Agent Consistency (2 Iterations))	8.0	1000	10.20
AOKVQA			
Baselines			
Perplexity of Direct Answer, Generated Confidence	N/a	1000	0.16
Self-Consistency based on Paraphrase	4	1000	3.90
DeCC Multi-Agent Consistency Stages			
Question Pre-Decomposition	3.2	1000	3.36
Sub-Question Answering and VLM Agent Reasoning	3.2	1000	0.54
LLM Agent Reasoning	3.2	1000	0.10
Second Question Decomposition	3.9	253	3.79
Second Sub-Question Answering and VLM Agent Reasoning	3.9	253	0.66
Second LLM Agent Reasoning	3.9	253	0.12
DeCC Expected Value (Multi-Agent Consistency)	$[1000 \times 3.2 + 253 \times 3.9]/1000 = 4.19$	1000	$[1000 \times (3.36 + 0.54 + 0.10) +$ $253 \times (3.79 + 0.66 + 0.12)]/1000 = 5.16$
DeCC Expected Value (VLM-Agent Consistency (2 Iterations))	7.1	1000	8.35
DeCC Expected Value (LLM-Agent Consistency (2 Iterations))	7.1	1000	8.57

Table 6: Computational costs analysis on VCR and AOKVQA.

ation and can then be reused to evaluate multiple VLMs, thus mitigating the impact on evaluation speed. The Sub-Question Answering and VLM Agent Reasoning + LLM Agent Reasoning takes 1.02 seconds per sample for VCR and 0.64 seconds per sample for AOKVQA. The second iteration of DeCC takes 5.22 seconds per sample for VCR and 4.57 seconds per sample for AOKVQA. However, the second decomposition is required only for samples where there is a disagreement between VLM Agent Consistency and LLM Agent Consistency from the first iteration, rather than for all samples. For example, only 25.3% of samples in AOKVQA and 36.6% of samples in VCR require the second iteration. Thus, the total time for the second iteration is significantly less than the first iteration. Overall the expected total time of DeCC Multi-Agent Setting taking into account both first and second round of decompositions is 6.89 seconds per sample for VCR and 5.16 seconds per sample for AOKVQA. Note that for VLM-Agent Consistency (2 Iterations) and LLM-Agent Consistency (2 Iterations), the expected total time is larger than the multi-agent setting because we need to do the second round of decomposition for all samples. With engineering optimizations and parallel processing, the computing time of DeCC can be further reduced. Note that DeCC does not require any training or tuning of thresholds while other baselines such as perplexity, generated numerical confidence scores, or self-consistency require tuning certain thresholds, necessitating annotated datasets. For example, we used 1,000 samples for threshold tuning in our experiments. Data annotation is usually time-consuming and costly, averaging 10 seconds per sample[7] with further filtering requests, and for specialized domains, this time can increase significantly, thus increasing the hidden costs. Compared with the data annotation, the time cost of DeCC is significantly lower. In conclusion, despite the additional computation introduced by DeCC compared to baselines, our comprehensive experimental evaluation validates the effectiveness and practicality of DeCC across various benchmarks and VLMs.

[7] CloudResearch: A Simple Formula for Predicting the Time to Complete a Study on Mechanical Turk

Algorithm 1 Multi-Agent Consistency Comparison over Task Decomposition for Reliability Measurement

```
Require: Question $Q$, Image $I$, Answer $A$, Decomposer, VLM for Evaluation, LLM for Reasoning
Ensure: Binary Reliability Score $\mathcal{R}$
Decomposer decomposes $Q$ into sub-questions
Generate sub-QA pairs by having VLM answer the sub-questions
Obtain $A^{'}_{V}$ and $A^{'}_{L}$ by reasoning over sub-QA pairs using VLM and LLM, respectively
if $A^{'}_{V}$ is consistent with $A$ then
    $Cons_{V'} \leftarrow 1$
else
    $Cons^{'}_{V} \leftarrow 0$
end if
if $A^{'}_{L}$ is consistent with $A$ then
    $Cons^{'}_{L} \leftarrow 1$
else
    $Cons^{'}_{L} \leftarrow 0$
end if
if $Cons^{'}_{V} = Cons^{'}_{L}$ then
    $\mathcal{R} \leftarrow Cons^{'}$                                    ▷ Direct determination
else
    Perform second-iteration decomposition and generate new sub-QA pairs
    Obtain $A^{''}_{V}$ and $A^{''}_{L}$ by reasoning over all sub-QA pairs using VLM and LLM, respectively
    if $A^{''}_{V}$ is consistent with $A$ then
        $Cons^{''}_{V} \leftarrow 1$
    else
        $Cons^{''}_{V} \leftarrow 0$
    end if
    if $A^{''}_{L}$ is consistent with $A$ then
        $Cons^{''}_{L} \leftarrow 1$
    else
        $Cons^{''}_{L} \leftarrow 0$
    end if
    if $Cons^{''}_{V} = Cons^{''}_{L}$ then
        $\mathcal{R} \leftarrow Cons^{''}$                 ▷ Direct determination after second iteration
    else
        if $Cons^{'}_{V} = Cons^{''}_{V}$ and $Cons^{'}_{L} = Cons^{''}_{L}$ then
            $\mathcal{R} \leftarrow Cons^{'}_{L}$                     ▷ LLM's consistency is used
        else if $Cons^{'}_{V} \neq Cons^{''}_{V}$ and $Cons^{'}_{L} \neq Cons^{''}_{L}$ then
            $\mathcal{R} \leftarrow Cons^{''}_{V}$                    ▷ VLM's consistency is used
        end if
    end if
end if
```

Few-Shot Prompt for Decomposition

Given an image and an associated main question, design pre-questions that focus on important contextual information in the image useful for answering the main question. Pre-questions should provide clues to answer the main question. Each pre-question should be short and easy to understand. Pre-questions should focus on context visual clues of the image. Pre-questions should provide clues to answer the main question.

Example scenario to illustrate the expected interaction pattern:
Main Question: Is this statement entailment, neutral or contradiction based on the image? Statement: ‘A professor is late to class’ Options: A: entailment, B: neutral, C: contradiction.
Pre-question 1: Is there a person in the image wearing clothing typically associated with a professor?
Pre-question 2: Is the person in the image displaying any behavior that could be interpreted as being late to class, such as being out of breath or looking at a clock?
Pre-question 3: Is there a classroom setting in the image, such as desks or a blackboard?

Example scenario to illustrate the expected interaction pattern:
Context: Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem. Main Question: Based on the arrows, which of the following organisms is a decomposer? Choices: A: mushroom, B: lichen
Pre-question 1: Does the mushroom eat any other organisms in the food web?
Pre-question 2: Does the lichen eat any other organisms in the food web?
Pre-question 3: Does the lichen produce any material that other organisms can use?
Pre-question 4: Does the mushroom produce any material that other organisms can use?
Pre-question 5: Does a decomposer produce any material that other organisms can use?

Example scenario to illustrate the expected interaction pattern:
Main Question: Is this statement entailment, neutral or contradiction based on the image? Statement: ‘Two children play in the park.’ Options: A: entailment, B: neutral, C: contradiction.
Pre-question 1: Are there any children in the image?
Pre-question 2: Are the two children playing in the park?

Example scenario to illustrate the expected interaction pattern:
User: Context: Use the graph to answer the question below. Main Question: Which month has the highest average precipitation in Santiago? Choices: A: March, B: October, C: June
Pre-question 1: What kind of graph is shown?
Pre-question 2: Does the graph show the average precipitation for each month in Santiago?
Pre-question 3: For which month is the bar highest in the graph?

Table 7: Few-Shot Prompt for Decomposition.

Few-Shot Prompt for Second-Iteration Decomposition

You will be given an image and an associated main question, and some sub-question-answer pairs. However, these sub-questions might not be sufficient to answer the main question due to lack of detail or conflicting answers. You need to design additional sub-questions that focus on important contextual information in the image useful for answering the main question. Each pre-question should be short, easy to understand, and provide clues to answer the main question.

Example scenario to illustrate the expected interaction pattern:
Main Question: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A professor is late to class’ Options: A: entailment, B: neutral, C: contradiction.
Sub-questions and answers:
Sub-question 1: Is there a person in the image wearing clothing typically associated with a professor?
Sub-answer 1: Yes.
Sub-question 2: Is the person in the image displaying any behavior that could be interpreted as being late to class, such as being out of breath or looking at a clock?
Sub-answer 2: No.
Sub-question 3: Is there a classroom setting in the image, such as desks or a blackboard?
Sub-answer 3: Yes.
Your return:
Additional Sub-question 1: What is the person’s age in the image?
Additional Sub-question 2: Is the person more likely to be a student or a professor?
Additional Sub-question 3: Is the person holding any books or papers?

Example scenario to illustrate the expected interaction pattern:
Context: Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem. Main Question: Based on the arrows, which of the following organisms is a decomposer? Choices: A: mushroom, B: lichen.
Sub-questions and answers:
Sub-question 1: Does the mushroom eat any other organisms in the food web?
Sub-answer 1: Yes.
Sub-question 2: Does the lichen eat any other organisms in the food web?
Sub-answer 2: No.
Sub-question 3: Does the lichen produce any material that other organisms can use?
Sub-answer 3: Yes.
Sub-question 4: Does the mushroom produce any material that other organisms can use?
Sub-answer 4: No.
Sub-question 5: Does a decomposer produce any material that other organisms can use?
Sub-answer 5: Yes.
Your return:
Additional Sub-question 1: Is there any arrow pointing towards the mushroom?
Additional Sub-question 2: Is there any arrow pointing towards the lichen?
Additional Sub-question 3: What is the mushroom’s role in the food web?
Additional Sub-question 4: What is the lichen’s role in the food web?

Table 8: Few-Shot Prompt for Second-Iteration Decomposition.

Few-Shot Prompt for Paraphrase

Your goal is to paraphrase the given question into 4 questions. Each question should only change the wording of the original question slightly or just replace a few words. The questions should be easy to understand and should not change the meaning of the original question. If the questions come with some choices, you should not change these choices.

Example scenario to illustrate the expected interaction pattern:
Main Question: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A professor is late to class’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 1: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A teacher is late to class’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 2: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A professor is tardy to class’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 3: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A professor is not on time for class’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 4: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘A teacher is not punctual for class’ Options: A: entailment, B: neutral, C: contradiction.

Example scenario to illustrate the expected interaction pattern:
Context: Below is a food web from a tundra ecosystem in Nunavut, a territory in Northern Canada. A food web models how the matter eaten by organisms moves through an ecosystem. The arrows in a food web represent how matter moves between organisms in an ecosystem. Main Question: Based on the arrows, which of the following organisms is a decomposer? Choices: A: mushroom, B: lichen
Paraphrased question 1: Based on the arrows, which of these choices is a decomposer? Choices: A: mushroom, B: lichen
Paraphrased question 2: Based on the arrows, which of the following is a decomposer? Choices: A: mushroom, B: lichen
Paraphrased question 3: Which of the following is a decomposer based on the arrows? Choices: A: mushroom, B: lichen
Paraphrased question 4: Which is a decomposer based on the figure? Choices: A: mushroom, B: lichen

Example scenario to illustrate the expected interaction pattern:
Main Question: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘Two children play in the park.’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 1: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘Two kids play in the park.’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 2: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘Two children are playing in the park.’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 3: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘Two kids are playing in the park.’ Options: A: entailment, B: neutral, C: contradiction.
Paraphrased question 4: Is this statement entailment, neutral, or contradiction based on the image? Statement: ‘There are two children playing in the park.’ Options: A: entailment, B: neutral, C: contradiction.

Example scenario to illustrate the expected interaction pattern:
User: Context: Use the graph to answer the question below. Main Question: Which month has the highest average precipitation in Santiago? Choices: A: March, B: October, C: June
Paraphrased question 1: Which month has the highest average rainfall in Santiago? Choices: A: March, B: October, C: June
Paraphrased question 2: Which month’s precipitation is the highest in Santiago? Choices: A: March, B: October, C: June
Paraphrased question 3: Which month has the most precipitation in Santiago? Choices: A: March, B: October, C: June
Paraphrased question 4: Which month has the most rainfall in Santiago? Choices: A: March, B: October, C: June

Note: Return the paraphrased questions. For each paraphrased question, you should return the entire set of choices as well.

Table 9: Few-Shot Prompt for Paraphrase.