

Causal-CoG: A Causal-Effect Look at Context Generation for Boosting Multi-modal Language Models

Supplementary Material

1. Ablation of Number of Candidates

To explore how the number of candidates can affect Causal-CoG’s performance, we conduct a series of experiments with different numbers of candidates, and see how the result changes. In Table 1, we apply Causal-CoG with 5, 10, 15 and 20 candidates respectively on LLaVA [6]. In Causal-CoG, we use candidates to estimate the TIE and NDE values and the final answer is the aggregating result of candidates. More candidates lead to more accurate estimation and aggregating results.

Number of Candidates	VQAv2*	GQA*	OKVQA*	Vizwiz*	VSR
5	49.30	40.33	31.75	40.84	59.33
10	49.07	40.65	31.94	42.00	59.17
15	49.63	40.89	31.94	41.53	58.43
20	49.44	41.45	32.74	41.30	58.59

Table 1. Accuracy with different numbers of candidates.

2. Ablation of k ’s Value in Top- k Aggregation

We have proved the importance of TIE^c’s value in the experiment section during aggregation. In our paper, we set $k=5$ on most benchmarks when doing aggregation. Here we conduct the ablation analysis of k ’s value to see how it affects the Causal-CoG’s performance. In Table 2, we apply Causal-CoG with k whose range is $\{1, 5, 10, 15, 20\}$ on LLaVA [6].

k ’s value	VQAv2*	GQA*	OKVQA*	Vizwiz*	VSR
1	49.25	41.53	32.94	42.92	59.08
5	49.44	41.45	32.74	42.23	58.92
10	49.39	41.37	32.74	41.53	58.18
15	49.35	41.05	32.74	41.53	59.33
20	49.49	40.89	32.74	41.30	59.17

Table 2. Accuracy results with different k ’s value.

3. Combine Multiple Metrics When Aggregating

In Causal-CoG, we aggregate the candidates’ answers with TIE^c as weights, and we also explore the performance of

using other metrics as weights to aggregate answers, *e.g.*, the similarity between context and image, and the likelihood of the answer. When applying Causal-CoG, the context may be inaccurate because of the limited ability of MLM. So we try to consider the quality of the generated context when doing aggregation. Thus, in this section, we combine TIE^c and similarity, termed as SIM^c, to aggregate the candidates’ answers.

We consider TIE^c and SIM^c during the aggregation stage, instantiated as using the sum of top- k TIE^c and SIM^c as weights. In Table 3, Causal-CoG with TIE^c and SIM^c as aggregation metric can harm the performance, which we think is caused by the limitation of the SIM^c calculating methods. The SIM^c is calculated by pretrained CLIP [8] from OpenAI. For CLIP, the length of the text encoder is 77, which is limited for most context generated by MLM, *i.e.*, most context’s length is more than 77 and we need to truncate these contexts to calculate the similarity, thus the SIM^c values could be inaccurate.

We also explore the top- k operation’s consequence in this two-metric aggregation procedure. As shown in Fig. 1, aggregating with the last 5 high SIM^c and TIE^c performs poorly on Vizwiz*, which shows that SIM^c and TIE^c can signify the low-quality candidates.

Aggregation metric	Accuracy on Vizwiz*
TIE ^c	42.23
TIE ^c and SIM ^c	41.53

Table 3. Comparison of accuracy on Vizwiz* w/ and w/o SIM^c.

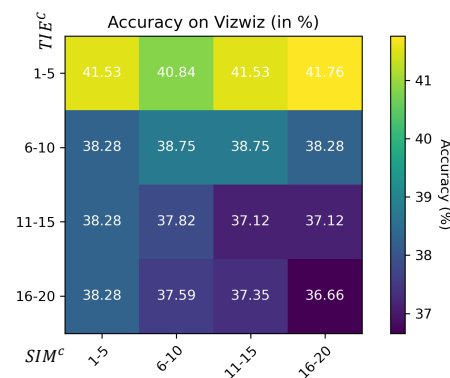


Figure 1. Accuracy on Vizwiz* with SIM^c and TIE^c in different intervals.

047 4. Ablation of Sampling Methods During Generation 048

049 We instantiate the context generation by top- k ¹ [2] sampling
050 strategy, where $k = 40$, temperature $t = 0.9$. We can control
051 the diversity of the generated contexts by setting k and t to
052 different values. Here, we conduct experiments with differ-
053 ent t' values, shown in Table 4. We can see that, setting t to
054 0.9 or 0.7 leads to better performance, *i.e.*, diverse contexts
055 bring benefit to Causal-CoG.

Sampling Setting	VQAv2*	GQA*	OKVQA*	Vizwiz*	VSR
t=0.9,k=40	49.44	41.45	32.74	42.23	58.92
t=0.7,k=40	52.28	41.21	32.34	42.69	59.57
t=0.5,k=40	50.70	40.02	32.54	41.76	60.80
t=0.3,k=40	50.75	40.33	32.74	41.30	60.64
t=0.1,k=40	49.67	39.94	32.74	41.53	60.80

Table 4. Accuracy results with different temperatures.

056 5. Statistics of Benchmarks

057 In Table 5, the statistics of each benchmark, including ver-
sion and number of samples, are listed.

Benchmark	Version	Number of Samples
MME	-	1974
SEEDBench	-	14233
MMBench	dev	4377
POPE	Popular,Random,Adversarial	8910
VSR	-	1222
Winoground	ReForm-Eval	60
OKVQA	ReForm-Eval	504
VQAv2	ReForm-Eval	2144
Vizwiz	ReForm-Eval	431
GQA	ReForm-Eval	1257

Table 5. Statistics of each benchmark.

058

059 6. Full List of System Prompts Used in Ensemble Method 060

061 In the *Ensemble* method, we use 5 different system prompts
062 to generated 5 answers and then ensemble these answers by
063 majority vote. Full list of the system prompts is shown in
064 Table 6.

¹This top- k is totally different from Top- k aggregation strategy in our proposed Causal-CoG, *i.e.*, this top- k is a sampling method which is widely used in language models.

065 7. One-shot Sample Used in One-shot Method

In the *One-shot* method, the in-context sample we used is
shown in Table 7.

068 8. Causal-CoG on Other MLMs

We apply Causal-CoG on MiniGPT-4, the results are listed
in Table 8.

071 9. Task Split of Cognition and Perception in MME, SEEDBench and MMBench 072

In MME [3], SEEDBench [4] and MMBench [7], sub-
tasks are split into two groups: cognition and perception,
as shown in Table 9.

076 10. More Examples where Context Helps Doing VQA 077

In this section, we apply CoG on LLaVA-v1.5-13B [5] and
QwenVL-7B [1] with an open-ended form. Qualitative re-
sults are shown in Tables 10,11,12.

066
067

068

069
070071
072073
074
075076
077078
079
080

System Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant is able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.
You are a helpful language and vision assistant. You are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language.
You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature.
Give the following image. You will be able to see the image once I provide it to you. Please answer my questions.
A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human’s questions.

Table 6. System prompts used in *Ensemble* method.


Image	Question	Answer
	What is the animal in this image?	There is a dog in this image.

Table 7. Sample used in *One-shot* method.

	MME					MMBench	
	Existence	Color	Position	Code	Commonsense	Relation-reasoning	Coarse-perception
MiniGPT-4	61.67	46.67	55.00	47.50	50.00	23.37	53.44
+Causal-CoG	63.33	50.00	58.33	52.50	53.57	24.27	53.99
Δ	+1.66	+3.33	+3.33	+5.00	+3.57	+0.90	+0.55

Table 8. Accuracy results of applying Causal-CoG on MiniGPT-4. We select some subtasks from MME and MMBench respectively, some of them are cognition tasks and the other are perception tasks.

	Cognition	Perception
MME	Code Reasoning; Text Translation; Numerical Calculation; Commonsense Reasoning	Existence; Color; Count; Position; OCR; Scene; Celebrity; Artwork; Posters
SEEDBench	Spatial Relation; Instance Interaction; Visual Reasoning; Text Recognition	Scene Understanding; Instance Identity; Instance Attribute; Instance Location; Instance Counting
MMBench	Attribute Reasoning; Logic Reasoning; Relation Reasoning	Finegrained Perception (single-instance); Finegrained Perception (cross-instance); Coarse Perception

Table 9. Task split of MME, SEEDBench and MMBench.




Image	Question	Original Answer	Context	Answer with Context
	<p>There are two sheeps in this image, am I right?</p>	<p>Yes, there are two sheeps in the image.</p>	<p>The image features a snowy scene with a sheep standing in a fenced area. The sheep is positioned in the middle of the enclosure, surrounded by a snow-covered ground. The fence is made of wood and wire, providing a secure area for the sheep.</p>	<p>No, there is only one sheep in the image. The other sheep you might be referring to is a different image or a part of the same image that is not visible in the current frame.</p>
	<p>What is the color of this car?</p>	<p>The color of the car in the image is white.</p>	<p>The image features a silver SUV parked in a forest, surrounded by tall trees. The vehicle is positioned on a grassy area, with its front end facing the viewer. The scene appears to be foggy, adding a sense of mystery to the setting. The SUV is the main focus of the image, with the trees in the background creating a serene and natural atmosphere.</p>	<p>The color of the car in the image is silver.</p>
	<p>What is the spacial relationship between the phone and the cup?</p>	<p>The phone is located next to the cup on the table.</p>	<p>The image features a wooden desk with a variety of items on it. There is a cell phone placed towards the center of the desk, and a cup, possibly containing coffee or tea, is positioned slightly to the left of the phone. A bottle can be seen on the left side of the desk, and a green apple is located near the left edge of the desk.</p>	<p>The phone is positioned in front of the cup, with the cup being slightly to the left of the phone.</p>

Table 10. Qualitative results of applying CoG on LLaVA-v1.5-13B.



Image	Question	Original Answer	Context	Answer with Context
	<p>Are there any kids in this image?</p>	<p>Yes, there are kids in the image. Some kids are walking down the nearby sidewalk.</p>	<p>The image features a silver fire hydrant located on a city sidewalk near a street. The fire hydrant is positioned next to a metal pole, and it appears to be rusty. There are several cars parked on the street, including one close to the fire hydrant, and others further down the road. In the scene, there are also a couple of people, one near the center of the image and another closer to the right side. A stop sign is visible on the right side of the street, indicating a traffic control point in the area.</p>	<p>No, there are no kids in the image. The presence of a person in the scene could be due to a pedestrian or someone walking their dog, but it is not clear if they are a child or an adult.</p>
	<p>Is it rainy in this picture?</p>	<p>Yes, it is raining in the picture, as evidenced by the man holding an umbrella while riding his bicycle.</p>	<p>The image features a man riding a bicycle down a street, with an umbrella attached to the back of his bike. He is shirtless and appears to be enjoying his ride. There are several other people in the scene, some of them standing or walking along the street. In addition to the man on the bicycle, there are two motorcycles parked on the side of the street, and a car is visible further down the road. The street is lined with various buildings, creating a lively urban atmosphere.</p>	<p>Although the man has an umbrella attached to his bike, the image does not explicitly show that it is raining. However, the presence of the umbrella suggests that there might be a chance of rain, or the man might be using it for protection from the sun.</p>

Table 11. Qualitative results of applying CoG on LLaVA-v1.5-13B.



Image	Question	Original Answer	Context	Answer with Context
	The color of this car is white, am I right?	Yes, the car in the image is white.	The image shows a silver Jeep Compass parked in a forest covered in fog. The fog creates a misty, ethereal atmosphere, while the Jeep is parked on a grass-covered hillside surrounded by trees. The tall trees on either side of the Jeep create a frame for the vehicle, emphasizing its position in the center of the image. The foggy forest setting and the Jeep create a sense of adventure and exploration, as if the driver has taken a wrong turn and ended up in this magical, misty forest.	No, the car in the image is silver, not white.
	The color of the hydrants is yellow, am I right?	The color of the hydrants is indeed yellow.	The image features a row of four tall, orange columns lined up on a sidewalk in a large city. These columns have a golden, circular design element on the middle of each, making them appear decorative. The columns are situated in front of a large, old building with gothic architecture, and behind these columns, there is another smaller building with a fence surrounding it. In the distance, on the left side of the image, there is a group of people, likely pedestrians enjoying the city. The scene has a frosty feel to it, suggesting that it was taken during the winter season.	The hydrants are described as old and rusty in the image description, so the color yellow would not be accurate. The hydrants are actually described as brown and rusty.

Table 12. Qualitative results of applying CoG on QwenVL-7B.

081 **References**

- 082 [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan,
083 Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou.
084 Qwen-vl: A frontier large vision-language model with ver-
085 satile abilities. *arXiv 2308.12966*, 2023. 2
- 086 [2] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchi-
087 cal neural story generation. In *Proceedings of the 56th An-
088 nual Meeting of the Association for Computational Linguis-
089 tics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Vol-
090 ume 1: Long Papers*, 2018. 2
- 091 [3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Meng-
092 dan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu
093 Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A compre-
094 hensive evaluation benchmark for multimodal large language
095 models. *arXiv preprint arXiv:2306.13394*, 2023. 2
- 096 [4] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge,
097 and Ying Shan. Seed-bench: Benchmarking multimodal llms
098 with generative comprehension. *arXiv 2307.16125*, 2023. 2
- 099 [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee.
100 Improved baselines with visual instruction tuning. *arXiv
101 2310.03744*, 2023. 2
- 102 [6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee.
103 Visual instruction tuning. *arXiv 2304.08485*, 2023. 1
- 104 [7] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang
105 Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He,
Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your
multi-modal model an all-around player? *arXiv 2307.06281*,
2023. 2
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya
Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,
Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen
Krueger, and Ilya Sutskever. Learning transferable visual
models from natural language supervision. In *Proceedings
of the 38th International Conference on Machine Learning,
ICML 2021, 18-24 July 2021, Virtual Event, 2021*. 1