# Chain-of-Description: What I can understand, I can put into words

**Anonymous ACL submission**

## Abstract

In this paper, we propose a novel strategy defined as Chain-of-Description (CoD) Prompting, tailored for Multi-Modal Large Language Models. This approach involves having the model first provide a detailed description of the multi-modal input before generating an answer to the question. When applied to models such as Qwen2-Audio, Qwen2-VL, and Qwen2.5-VL, CoD Prompting significantly enhances performance compared to standard prompting methods. This is demonstrated by nearly a 4% improvement in the speech category of the audio benchmark AIR-Bench-Chat and a 5.3% improvement in the hard-level portion of the vision benchmark MMMU_Pro. Our ablation study further validates the effectiveness of CoD Prompting.

## 1 Introduction

Multi-Modal Large Language Models (Wu et al., 2024; Yin et al., 2023; Chen et al., 2024a) (MLLMs), which encompass Large Audio-Language Models (Xie and Wu, 2024; Das et al., 2024; Fang et al., 2024; Wu et al., 2023a; Chu et al., 2023, 2024) (LALMs) and Large Vision-Language Models (Chen et al., 2023, 2024b; Li et al., 2024a; Shi et al., 2024; Wang et al., 2024; Team, 2025) (LVLMs), have shown considerable potential in managing a variety of input types. These models are generally based on Large Language Models (LLMs) and employ audio/vision encoders to align multi-modal inputs with text. However, the conventional method of directly generating answers from audio or visual inputs might not fully capitalize on the information and comprehension that the models can potentially extract. Given that the training paradigm for most MLLMs involves aligning multi-modal inputs with text, we explored whether there is an inference strategy that can not only explicitly align these inputs but also improve the quality of the results produced by MLLMs.

We believe that *"What I can understand, I can put into words."* This implies that if a model can generate a detailed description of the input, it indicates a deeper level of understanding. Based on this idea, we proposed the Chain-of-Description (CoD) prompting for MLLMs, which involves having the model first provide a detailed description of the multi-modal input before answering the question.

We have implemented CoD Prompting on LALMs and LVLMs, specifically with the Qwen2-Audio (Chu et al., 2024), Qwen2-VL (Wang et al., 2024), and Qwen2.5-VL (Team, 2025) models. Through experiments conducted on their respective test sets, we have demonstrated that CoD significantly enhances the reasoning performance of these models compared to the standard approach. Specifically, in the speech testset AIR-Bench-Chat (Yang et al., 2024), the Qwen2-Audio model performance for human speech improved by nearly 4%. In the image testset MMLU_Pro (Yue et al., 2024), for the hard-level category, the performance of Qwen*-VL increased by 5.3%.

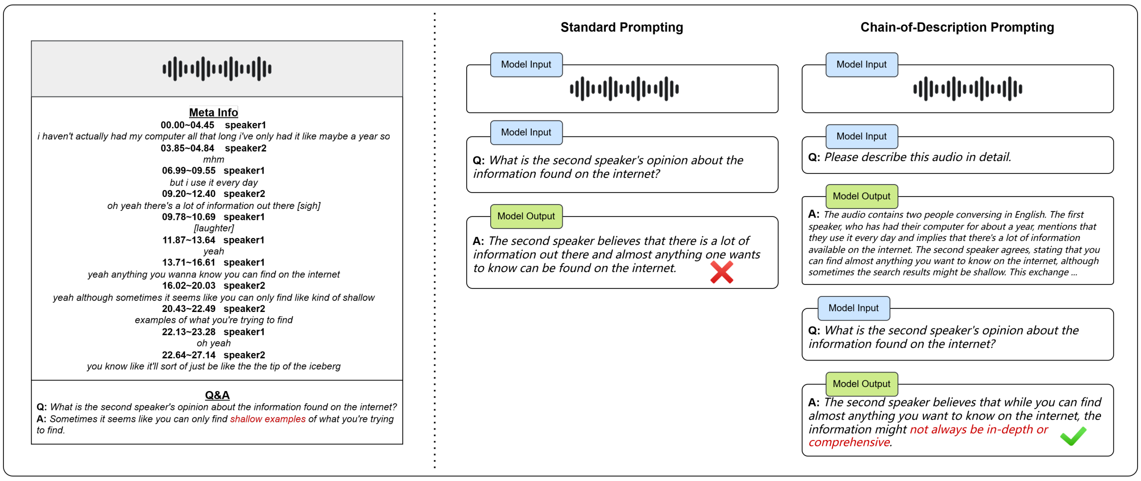
Furthermore, in our experiments with LALMs, we analyzed information density to substantiate why CoD is effective. In our experiments with LVLMs, we verified the correctness of CoD by constructing better descriptions.

## 2 Chain-of-Description Prompting

**Motivation:** *What I can understand, I can put into words.*

The motivation of our proposed Chain-of-Description (CoD) Prompting is if a model can generate a detailed description of the input, it indicates a deeper level of understanding. CoD Prompting can be described as follows:

1. Supply audio/vision inputs to MLLLMs and ask the models produce detailed descriptions.
2. MLLMs generate a comprehensive textual representation.



**Meta Info**
**00.00~04.45 speaker1**
*i haven't actually had my computer all that long i've only had it like maybe a year so*
**03.85~04.84 speaker2**
*mhm*
**06.99~09.55 speaker1**
*but i use it every day*
**09.20~12.40 speaker2**
*oh yeah there's a lot of information out there [sigh]*
**09.78~10.69 speaker1**
*[laughter]*
**11.87~13.64 speaker1**
*yeah*
**13.71~16.61 speaker1**
*yeah anything you wanna know you can find on the internet*
**16.02~20.03 speaker2**
*yeah although sometimes it seems like you can only find like kind of shallow*
**20.43~22.49 speaker2**
*examples of what you're trying to find*
**22.13~23.28 speaker1**
*oh yeah*
**22.64~27.14 speaker2**
*you know like it'll sort of just be like the the tip of the iceberg*

**Q&A**
**Q:** *What is the second speaker's opinion about the information found on the internet?*
**A:** *Sometimes it seems like you can only find shallow examples of what you're trying to find.*

Standard Prompting:

**Q:** *What is the second speaker's opinion about the information found on the internet?*

**A:** *The second speaker believes that there is a lot of information out there and almost anything one wants to know can be found on the internet.* ✗

Chain-of-Description Prompting:

**Q:** *Please describe this audio in detail.*

**A:** *The audio contains two people conversing in English. The first speaker, who has had their computer for about a year, mentions that they use it every day and implies that there's a lot of information available on the internet. The second speaker agrees, stating that you can find almost anything you want to know on the internet, although sometimes the search results might be shallow. This exchange ...*

**Q:** *What is the second speaker's opinion about the information found on the internet?*

**A:** *The second speaker believes that while you can find almost anything you want to know on the internet, the information might not always be in-depth or comprehensive.* ✓

Figure 1: An example of using Standard Prompting and our Chain-of-Description (CoD) Prompting for Large Audio-Language Models (LALMs).

3. Thereafter, introduce the queries into MLLMs.
4. MLLMs generate responses pertinent to the queries.

For LALMs, describing speech context, background sounds, and other audio features comprehensively helps the model better understand audio inputs. Figure 1 illustrates examples of using Standard Prompting and CoD Prompting for LALMs. Likewise, for LVLMs, refer to Figure 2 in Appendix A, detailing objects, scenes, colors, and spatial relationships in images enhances comprehension. Focusing on the description process first aims to establish a strong foundation for generating higher-quality answers, improving MLLMs' overall performance.

## 3 Experiments with CoD in LALMs

### 3.1 Experimental Setup

**Model** The model used for our experiments is Qwen2-Audio[1], a state-of-the-art open-sourced LALM capable of processing various audio inputs and generating textual responses.

**Evaluation Dataset** The dataset we utilized is the AIR-Bench[2], which is the first and widely adopted benchmark designed to assess the comprehension capabilities of LALMs across various audio signals, including human speech, natural sounds, and music. Following prior research, we conducted detailed evaluations on all four subcategories of the AIR-Bench Chat Benchmark (AIR-Bench-Chat): Speech, Sound, Music, and Mixed Audio. More see Appendix B.

**Evaluation Method** Building upon previous work (Chu et al., 2024; Yang et al., 2024), we employed an evaluation method that utilizes a LLM as the judge (Li et al., 2024b; Gu et al., 2024). Specifically, we utilized a LLM to rate both the ground truth answer and the model prediction on a scale of 1 to 10. The final score is the average of these ratings. Considering cost-effectiveness, we chose gpt-4o-mini as our evaluation LLM.

In practice, we rated the model predictions from both Standard Prompting and CoD Prompting against the ground truth answer. This process yielded two sets of scores for the ground truth answer, which may exhibit minor differences. To facilitate more effective comparison, we calculated the ratio $r$ of the model prediction $p$ score $s_p$ to the ground truth answer $gt$ score $s_{gt}$. The $r$ quantifies the alignment between the prediction and the ground truth. See Appendix C.

### 3.2 Results

Based on the evaluation results presented in Table 1, it is evident that the Qwen2-Audio model's alignment with the ground truth answers has improved across all subcategories after adopting the CoD Prompting. Particularly in the Speech category, the alignment reached 95.02%, marking an increase of nearly 4% compared to the Standard Prompting. Other categories such as Sound, Music, and Mixed also experienced an enhancement of about 1% each. On average across all categories, there was an increase of 1.79%. This indicates that

---

[1] https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct

[2] https://github.com/OFA-Sys/AIR-Bench

| | Speech | Sound | Music | Mixed |
|---|---|---|---|---|
| | | *Standard Prompting* | | |
| $s_{gt}$ | 8.23 | 8.01 | 7.96 | 8.38 |
| $s_p$ | 7.51 | 6.93 | 6.87 | 6.49 |
| $r$ | 91.24% | 86.48% | 86.31% | 77.41% |
| | | *CoD Prompting* | | |
| $s_{gt}$ | 8.04 | 7.77 | 8.12 | 8.12 |
| $s_p$ | 7.64 | 6.83 | 7.08 | 6.38 |
| $r$ | **95.02%** | **87.87%** | **87.22%** | **78.50%** |

Table 1: The evaluation results for LALMs. These results are based on the AIR-Bench-Chat dataset among four subcategories: Speech, Sound, Music, and Mixed, where the *gpt-4o-mini* rated the model predictions ($p$) of the *Qwen2-Audio* model under both Standard Prompting and CoD Prompting and ground truth answers ($gt$) on a scale of 1 to 10. The average scores for the ground truth answer and model prediction are denoted as $s_{gt}$ and $s_p$. The alignment between the prediction and the ground truth is measured by $r = \frac{s_p}{s_{gt}}$.

the CoD Prompting method significantly enhances the consistency of the model's predictions with the ground truth answers. More in Appendix D.

### 3.3 Ablation Study

**Question:** *Why does CoD Prompting perform better in the Speech category?*

We analyze information density to explain why CoD Prompting performs better in the Speech category. We can consider the description as a textual representation of the audio, where the quantity of description indicates the level of information density. Compared to Sounds and Music, human speech offers a richer array of information, including textual content, emotional expressions, and background noise.

| | Speech | Sound | Music |
|---|---|---|---|
| $\Delta r$ | 3.78% | 1.40% | 0.91% |
| $id$ | 3.91 | 1.30 | 2.52 |

Table 2: The analysis of information density focuses on the Speech, Sound, and Music categories. In this analysis, the information density is represented by the number of tokens per second in the description, which is defined as $id$. $\Delta r$ represent the $r$ improvements of CoD Prompting compared to Standard Prompting.

We conducted a statistical analysis of the audio in Speech, Sound, and Music categories, calculating the average length of the description generated per second, which is the number of tokens in the description divided by the duration of the audio, defined as $id$. We did not include the Mixed category in our statistics as it is a combination of the other three categories.

The results from Table 2 indicate that the Speech category can generate an average of nearly 4 tokens of description per second, significantly higher than the Sound and Music categories. Therefore, the use of CoD Prompting in the Speech category yields better results. Although Music has a higher information density than Sound, the improvement when using CoD Prompting in Music and Sound categories is roughly the same, a phenomenon that requires further analysis.

## 4 Experiments with CoD in LVLMs

### 4.1 Experimental Setup

**Model** Our experiments utilized models from the Qwen-VL series, including Qwen2-VL-7B-Instruct[3], Qwen2.5-VL-7B-Instruct[4] and Qwen2.5-VL-72B-Instruct[5].

**Evaluation Dataset** In our experiments, we utilized the MMMU_Pro[6] dataset. MMMU_Pro is an enhanced multimodal benchmark designed to rigorously assess the true understanding capabilities of LVLMs. Specifically, we employed the standard validation data within MMMU_Pro that has been enhanced with 10 options. Additionally, this set can be categorized into three levels of difficulty: Easy, Medium, and Hard. More see Appendix E.

**Evaluation Method** As MMMU_Pro is a multiple-choice dataset, we can directly calculate the accuracy of model predictions. We use both Standard Prompting and CoD Prompting methods to generate the answer options.

### 4.2 Results

Table 3 presents the evaluation results for LVLMs. Among the 7B-sized models, the results indicate that the Qwen2.5-VL-7B-Instruct outperforms the Qwen2-VL-7B-Instruct across all difficulty levels. Surprisingly, CoD Prompting did not demonstrate effectiveness at all levels; instead, it showed significant improvement at the Hard level, increasing the accuracy of both models by 5.3%. This suggests that CoD Prompting is more beneficial for more

---

[3] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct
[4] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
[5] https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
[6] https://huggingface.co/datasets/MMMU/MMMU_Pro

|  | Easy | Medium | Hard |
|---|---|---|---|
| *Qwen2-VL-7B-Instruct* |  |  |  |
| Standard | 40.91% | 24.91% | 16.67% |
| CoD | 39.77% | 22.30% | **21.97%** |
| *Qwen2.5-VL-7B-Instruct* |  |  |  |
| Standard | 40.91% | 26.77% | 16.67% |
| CoD | **42.61%** | 26.02% | **21.97%** |
| *Qwen2.5-VL-72B-Instruct* |  |  |  |
| Standard | 40.91% | 26.02% | 18.18% |
| CoD | **43.18%** | **26.02%** | **23.48%** |

Table 3: The evaluation results for LVLMs. We utilized the Qwen2-VL-7B-Instruct and Qwen2.5-VL-7B-Instruct models to conduct tests on the standard 10-option validation set of the MMMU_Pro dataset. The tests were carried out using both Standard Prompting and CoD Prompting approaches. For simplicity in the results table, we refer to these methods as "Standard" and "CoD" respectively.

complex images or questions. A detailed analysis of the Easy and Medium level cases revealed that due to the huge information density in images, the extensive descriptions generated did not necessarily cover the key points of the questions, potentially leading to a negative impact on responses. Moreover, this negative effect was less pronounced in the latest Qwen2.5-VL-7B-Instruct model.

The Qwen2.5-VL-72B-Instruct outperformed the Qwen2.5-VL-7B-Instruct across all difficulty levels, aligning with the well-known conclusion that larger model sizes lead to stronger performance. Our CoD Prompting method also achieved consistent improvements on the Qwen2.5-VL-72B-Instruct, with a 5.3% increase at the Hard level. Additionally, it maintained either no decline or a slight improvement at the Easy and Medium levels.

### 4.3 Ablation Study

**Question:** *How would the performance be affected if the model could generate higher-quality descriptions?*

The key to our CoD Prompting method lies in generating high-quality and accurate descriptions, which can lead to improved model performance. Knowing that the Qwen2.5-VL-7B-Instruct model outperforms the Qwen2-VL-7B-Instruct, we hypothesize that the descriptions generated by Qwen2.5-VL-7B-Instruct would be superior.

We conducted experiments on the Qwen2-VL-7B-Instruct model using descriptions generated by Qwen2.5-VL-7B-Instruct. As shown in Table 4, the results across all difficulty levels demonstrated positive improvements, confirming that higher quality descriptions can yield better outcomes.

|  | Easy | Medium | Hard |
|---|---|---|---|
| Standard | 40.91% | 24.91% | 16.67% |
| CoD | 39.77% | 22.30% | **21.97%** |
| CoD* | **42.61%** | **25.65%** | **21.97%** |

Table 4: The ablation study for Qwen2-VL-7B-Instruct. We compared two scenarios: in the CoD approach, the description is generated by the model itself; whereas in the CoD* approach, the description is produced by the latest and stronger model Qwen2.5-VL-7B-Instruct.

## 5 Related Work

We outline some related work and explain how our approach differs from these efforts.

Chain-of-Thought (Wei et al., 2022) (CoT) primarily aims to enhance the reasoning capabilities of text LLMs by breaking down problems into step-by-step solutions. Our method is inspired by CoT, but our CoD is not a variant of CoT; it is a new strategy specifically designed for multi-modal inputs. In theory, CoD and CoT could be combined to bolster the performance of MLLMs.

Wu et al. (2023b) focuses only on LVLMs and also mentions descriptions. However, their descriptions are related directly to the questions. In contrast, we do focus not on specific questions but explore a more general scenario and verify its effectiveness. Zhang et al. (2024) also concentrates on LVLMs, but their study emphasizes enhancing reasoning tasks, proposing rationale generation before answer inference. Our work is not limited to reasoning tasks but targets more general scenarios. Vijayaraghavan et al. (2024) introduced a method with the same name as ours, but their research direction is about improving the performance of Code LLM.

## 6 Conclusion

This study introduces Chain-of-Description (CoD) Prompting, which significantly enhances the reasoning capabilities of MLLMs. Experiments conducted on models such as Qwen2-Audio, Qwen2-VL, and Qwen2.5-VL demonstrate the effectiveness of CoD Prompting in improving model comprehension and response accuracy for multi-modal inputs, offering a promising direction for future research.

## 7 Limitations

Although our experimental results have been positive, the sheer number of open-source MLLMs and benchmark datasets prevents us from validating each one. Furthermore, based on our motivation, we believe that extensive multi-modal description training during the pre-training phase of MLLMs could yield significant benefits. Regrettably, this type of experiment is particularly resource-intensive, making it infeasible for us to complete swiftly.

## References

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Jun Yao, Lanqing Hong, Lu Hou, and Hang Xu. 2024a. EMOVA: empowering language models to see, hear and speak with vivid emotions. *CoRR*, abs/2409.18042.

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, Hewei 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. *CoRR*, abs/2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *CoRR*, abs/2312.14238.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J. Han, and Katrin Kirchhoff. 2024. Speechverse: A large-scale generalizable audio language model. *CoRR*, abs/2405.08295.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *CoRR*, abs/2409.06666.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *CoRR*, abs/2412.05579.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. 2024. Lmfusion: Adapting pretrained language models for multimodal generation. *CoRR*, abs/2412.15188.

Qwen Team. 2025. Qwen2.5-vl.

Prashanth Vijayaraghavan, Apoorva Nitsure, Charles Mackin, Luyao Shi, Stefano Ambrogio, Arvind Haran, Viresh Paruthi, Ali El-Zein, Dan Coops, David Beymer, Tyler Baldwin, and Ehsan Degan. 2024. Chain-of-descriptions: Improving code llms for VHDL code generation and summarization. In *Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD, MLCAD 2024, Salt Lake City, UT, USA, September 9-11, 2024*, pages 28:1–28:10. ACM.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. 2023a. On decoder-only architecture for speech-to-text and large language model integration. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. 2023b. The role of chain-of-thought in complex vision-language reasoning task. *CoRR*, abs/2311.09193.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *CoRR*, abs/2408.16725.

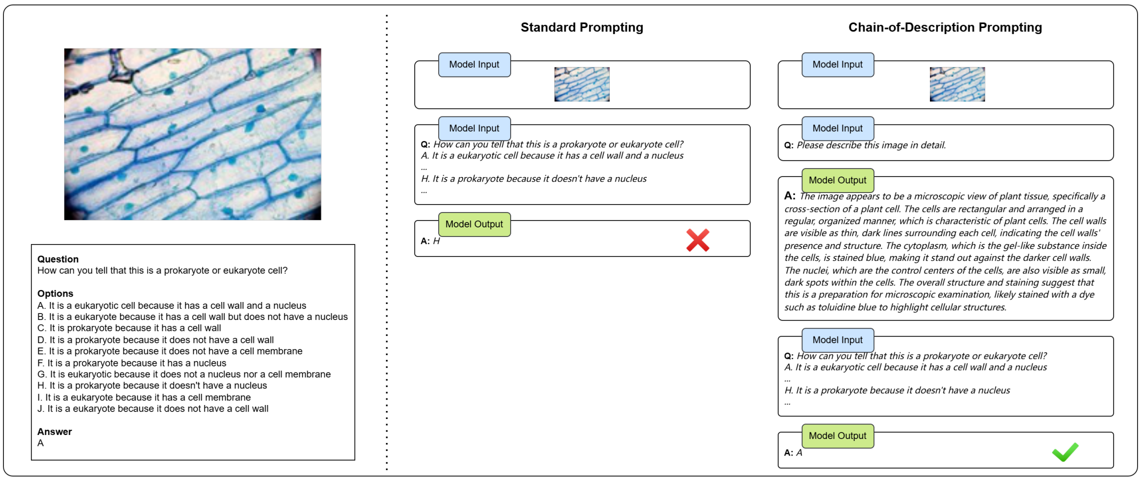
Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1979–1998. Association for Computational Linguistics.

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *CoRR*, abs/2409.02813.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

# A Chain-of-Description Prompting



**Question**
How can you tell that this is a prokaryote or eukaryote cell?

**Options**
A. It is a eukaryotic cell because it has a cell wall and a nucleus
B. It is a eukaryote because it has a cell wall but does not have a nucleus
C. It is prokaryote because it has a cell wall
D. It is a prokaryote because it does not have a cell wall
E. It is a prokaryote because it does not have a cell membrane
F. It is a prokaryote because it has a nucleus
G. It is eukaryotic because it does not a nucleus nor a cell membrane
H. It is a prokaryote because it doesn't have a nucleus
I. It is a eukaryote because it has a cell membrane
J. It is a eukaryote because it does not have a cell wall

**Answer**
A

**Standard Prompting**

**Q:** *How can you tell that this is a prokaryote or eukaryote cell?*
*A. It is a eukaryotic cell because it has a cell wall and a nucleus*
...
*H. It is a prokaryote because it doesn't have a nucleus*
...

**A:** *H*

**Chain-of-Description Prompting**

**Q:** *Please describe this image in detail.*

**A:** *The image appears to be a microscopic view of plant tissue, specifically a cross-section of a plant cell. The cells are rectangular and arranged in a regular, organized manner, which is characteristic of plant cells. The cell walls are visible as thin, dark lines surrounding each cell, indicating the cell walls' presence and structure. The cytoplasm, which is the gel-like substance inside the cells, is stained blue, making it stand out against the darker cell walls. The nuclei, which are the control centers of the cells, are also visible as small, dark spots within the cells. The overall structure and staining suggest that this is a preparation for microscopic examination, likely stained with a dye such as toluidine blue to highlight cellular structures.*

**Q:** *How can you tell that this is a prokaryote or eukaryote cell?*
*A. It is a eukaryotic cell because it has a cell wall and a nucleus*
...
*H. It is a prokaryote because it doesn't have a nucleus*
...

**A:** *A*

Figure 2: An example of using Standard Prompting and our Chain-of-Description (CoD) Prompting for Large Vision-Language Models (LVLMs).

Figure 2 shows examples of using Standard Prompting and Chain-of-Description Prompting for Large Vision-Language Models.

## B Statistic and Cases for experiments in LALMs

AIR-Bench-Chat contains 2,200 samples.

**Question**

How did the first person react to the second person's situation with restaurants?

**Answer**

The first person reacted by saying 'wow yeah'.

**Standard Out**

The first person expressed sympathy but also indicated that they do not smoke and are happy about it, suggesting a disconnect from the second person's experience.

**CoD Out**

The first person reacted by saying 'wow yeah,' which could indicate surprise or empathy with the second person's situation.

**Cod Description**

The audio contains two speakers discussing the topic of smoking and its impact on their rights and choices. The first speaker, who is not a smoker, expresses that they feel it infringes on non-smokers' rights but is happy about the situation since they do not smoke. The second speaker, also not a smoker, lives with smokers and does not believe that smoking should be banned from every place. They mention that they cannot go out to eat at certain restaurants because of smoking policies. The conversation ends with both speakers expressing disagreement with the idea of complete smoking bans.

Table 5: Case 1 for experiments in LALMs

**Question**

What kind of weather is mentioned as being present?

**Answer**

Rain is mentioned.

**Standard Out**

Colder and damper weather.

**CoD Out**

Colder weather and rain are mentioned.

**Cod Description**

The audio contains two people engaging in conversation, discussing the weather and their involvement in a research study. One individual mentions that it is colder where they are and implies that the weather could be related to the study. The other person responds by mentioning that it is damp and rainy where they are, which seems to align with the first person's comment about the cold. Further into the conversation, one of them asks if the other is a college student, to which the other responds affirmatively but clarifies that they have never been linked to a participant before. They both seem surprised and excited about this experience being their first time participating in such a study.

Table 6: Case 2 for experiments in LALMs

Table 5 and 6 are two cases for experiments in LALMs.

## C Evaluation Prompts of gpt-4o-mini for experiments in LALMs

---

You are a helpful and precise assistant for checking the quality of the answer.
[Detailed Audio Description]
XAudioX
[Question]
XQuestionX
[The Start of Assistant 1s Answer]
XAssistant1X
[The End of Assistant 1s Answer]
[The Start of Assistant 2s Answer]
XAssistant2X
[The End of Assistant 2s Answer]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question and audio description displayed above. AI assistants are provided with detailed audio descriptions and questions.
Please rate the helpfulness, relevance, accuracy, and comprehensiveness of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space.

---

Table 7: Evaluation Prompts of gpt-4o-mini for experiments in LALMs

Table 7 shows the evaluation prompts of gpt-4o-mini for experiments in LALMs. "Assistant 1" and "Assistant 2" correspond to the ground truth answer and the model prediction, respectively. To further enhance the fairness of the evaluation, we also swapped the positions of the ground truth answer and the model prediction, then recalculated the final score.

## D Evaluation Results for experiments in LALMs

| | Speech | | | Sound | | | Music | | | Mixed | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | No-S | Swap | Avg | No-S | Swap | Avg | No-S | Swap | Avg | No-S | Swap | Avg |
| *Standard Prompting* | | | | | | | | | | | | |
| $s_{gt}$ | 8.23 | 8.24 | 8.23 | 7.91 | 8.10 | 8.01 | 8.13 | 7.79 | 7.96 | 8.42 | 8.34 | 8.38 |
| $s_{gt}$ | 7.63 | 7.39 | 7.51 | 7.03 | 6.82 | 6.93 | 6.95 | 6.79 | 6.87 | 6.50 | 6.48 | 6.49 |
| $r(\%)$ | 92.80 | 89.69 | 91.24 | 88.80 | 84.20 | 86.48 | 85.54 | 87.11 | 86.31 | 77.14 | 77.68 | 77.41 |
| *CoD Prompting* | | | | | | | | | | | | |
| $s_{gt}$ | 8.11 | 7.98 | 8.04 | 7.81 | 7.73 | 7.77 | 8.22 | 8.02 | 8.12 | 8.24 | 8.00 | 8.12 |
| $s_p$ | 7.81 | 7.47 | 7.64 | 7.01 | 6.64 | 6.83 | 7.17 | 6.99 | 7.08 | 6.40 | 6.36 | 6.38 |
| $r(\%)$ | 96.36 | 93.65 | 95.02 | 89.78 | 85.95 | 87.87 | 87.29 | 87.15 | 87.22 | 77.62 | 79.41 | 78.50 |

Table 8: All evaluation results for experiments in LALMs. "No-S" indicates that "Assistant 1" is the ground truth answer and "Assistant 2" is the model prediction in the evaluation prompts. "Swap" signifies that in the evaluation prompts, "Assistant 2" is the ground truth answer and "Assistant 1" is the model prediction. "Avg" represents the average score between "No-S" and "Swap".

Table 8 shows the all evaluation results for experiments in LALMs.

## E Statistic and Cases for experiments in LVLMs

MMLU_Pro contains 577 validation samples.

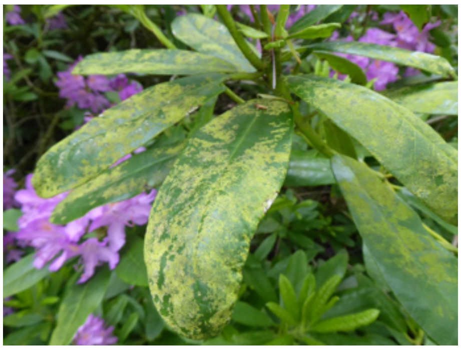


Figure 3: A case image.

**Question**

What is the substance that is developing on these leaves?

**Options**

A. Don’t know and don’t want to guess B. Powdery mildew C. Moss D. Rust E. Lichen F. Sooty mould G. Bacterial leaf spot H. Fungus I. Downy mildew J. Algae

**Answer**

J.

**Standard Out**

B.

**CoD Out**

J.

**Cod Description**

The image depicts a close-up view of a plant with green leaves that are covered in a yellowish-green substance, possibly algae or a fungal growth. The leaves are broad and elongated, with visible veins running through them. The plant appears to be healthy overall, with vibrant green foliage and a dense arrangement of leaves. In the background, there are clusters of purple flowers, which are likely part of the same plant or a nearby species. The flowers have multiple petals and are arranged in a dense, bushy manner. The overall scene suggests a natural, outdoor setting, possibly a garden or a forested area. The combination of the green leaves and purple flowers creates a visually appealing contrast.

Table 9: Cases for experiments in LALMs

Table 9 are cases of Figure 3 for experiments in LALMs.