From Head to Tail: Towards Balanced Representation in Large Vision-Language Models through Adaptive Data Calibration

Supplementary Material

A. Details of Experiments

A.1. Benchmarks

All benchmarks we used and their abbreviations are introduced as follows.

- **VQA$^{\mathbf{v2}}$**: The Visual Question Answering v2 dataset [5] consists of 265,016 images, each with 5.4 questions on average, requiring vision, language, and commonsense understanding, with 10 ground-truth and 3 plausible but incorrect answers for evaluation.
- **VQA$^{\mathbf{T}}$**: The TextVQA dataset [16] includes 45,336 questions over 28,408 OpenImages images, requiring models to read and reason about text within images for answers.
- **VQA$^{\mathbf{OK}}$**: The Open-Ended Knowledge Visual Question Answering dataset [14] includes over 14,000 questions that require integrating visual content with external knowledge, such as Wikipedia, for final accurate answers.
- **GQA**: GQA [6] is a large-scale dataset comprising over 22 million questions generated from scene graphs of 113,000 images. It is specifically designed to assess models on visual reasoning and compositional question answering, with a focus on reducing language biases.
- **SQA$^{\mathbf{I}}$**: ScienceQA-IMG [13] is a multimodal dataset comprising 21,208 science questions, each accompanied by corresponding images and explanations. It is designed to evaluate models' capabilities in answering science-related questions through multimodal reasoning.
- **POPE**: The Polling-based Object Probing benchmark [9] evaluates vision-language models' ability to detect hallucination by prompting them with classification questions regarding the presence of specific objects in an image.
- **SEED**: The SEED Bench [8] is a large-scale benchmark with 19,000 multiple-choice questions across 12 dimensions, designed for efficient evaluation of LVLMs without human intervention.
- **SEED$^{\mathbf{2}}$**: SEED Bench v2 [7] is a comprehensive benchmark with 24,000 multiple-choice questions across 27 dimensions, comprehensively evaluating text and image generation capabilities of LVLMs.
- **MMMU**: The Massive Multi-discipline Multimodal Understanding benchmark [18] is designed to evaluate multimodal models on complex, college-level tasks that require subject-specific knowledge and advanced reasoning.
- **MME$^{\mathbf{P}}$**: The Multimodal Evaluation Benchmark [4] assesses LVLMs' perception and cognition through 14 subtasks, including object recognition and reasoning. This paper focuses on its perception subset.
- **MMB$^{\mathbf{CN}}$**: MMBench [12] is a benchmark with 3,000 multiple-choice questions across 20 dimensions, assessing vision-language models' perceptual and cognitive abilities. CN denotes its Chinese validation set.
- **MMB**: The English validation subset of MMBench [12];
- **MMS**: MMStar [2] is a benchmark with 1,500 samples, assessing six core capabilities across 18 axes to evaluate LVLMs' visual comprehension in complex scenarios.
- **QB$^{\mathbf{2}}$**: Q-Bench 2 [17] is a benchmark for evaluating multimodal models on low-level vision tasks, focusing on visual perception, description, and quality assessment with datasets like LLVisionQA and LLDescribe.

A.2. Detailed Results of Ablation Study

We conducted an ablation study on different balancing combinations and synthesis methods. In the ablation study of different rebalancing combinations, we conduct the DR stage using different combinations of four perspectives, i.e., one or more from (Token, Object, Co-occurrence, and Interrogation) to validate the effectiveness of different perspectives. The detailed results of the balancing ablation experiment are presented in Table 2. Although some checkpoints achieved similar average results, we found that combining all perspectives yields the best performance in terms of both the number of top results and performance stability.

Additionally, we conducted an ablation study on different synthesis methods. The results of the augmentation and synthesis experiments are presented in Table 3. Obviously, synthesizing from **ALL** perspectives (as outlined in Section 4.2.2) yields the best performance.

A.3. Detailed Results of Main Experiment

Beyond the main experiments, we conduct pure data augmentation on the original instruction-tuning dataset of LLaVA 1.5, focusing solely on the DS stage applied to the original training data. The resulting augmented data is used to instruction-tune LLaVA, which is then evaluated on various benchmarks. As shown in Table 1, our ADR framework consistently surpasses most pure augmentation checkpoints on the majority of benchmarks, with a few exceptions, such as MMMU, MMB, and VQA$^{\text{v2}}$.

A.4. Qualitive Results

We present the full qualitative results in Figure 1. LLaVA 1.5 often fails to provide accurate responses when addressing tail questions. However, with the integration of our ADR framework, the model demonstrates significant improvement in recognizing and handling tail concepts. Additionally, we showcase more examples of our synthesized data in Figure 2. This synthesis process enriches the tail data with additional instances, effectively boosting the model's generalization and performance in underrepresented scenarios.

B. Details of Analyzing Stage

B.1. Examples of Entities

Different kinds of entities are extracted from four perspectives: Token, Object, Co-occurrence, and Interrogations. The top 20 frequently-shown entities from instruction-tuning data of LLaVA 1.5 are displayed in Figure 3.

B.2. Implement Details of Analyzing Stage

In this work, we construct the entity distribution using both the pretraining and instruction-tuning datasets from LLaVA 1.5, specifically LCS558K and Instructmix665K. To compare the differences between training and test data further, we also incorporate portions of the distributions from POPE and MME within the same figure. The complete results are presented in Figure 4. As illustrated, all pretraining, instruction-tuning, and evaluation datasets exhibit LT issues. However, the frequency distributions of training and evaluation data differ significantly.

In the Analyzing stage, token entities are extracted using Stanza[1] [15] as the POS parser. For object entities, we initially use LLaMA 3 70B Instruct[2] [3] to detect potential object-related vocabulary, followed by GroundingDINO[3] [11] to extract actual objects from the image. For co-occurrence distribution construction, we use Neo4j[4] to create an undirected graph. To construct interrogation entity distributions, we utilize LLaMA 3 70B Instruct[2] [3] to extract interrogation words.

B.3. Analysis of Failed Cases

We experiment to observe the distribution location of failed cases. We first extract all entities within the failed cases and calculate the max, min, and average location of these entities in the pertaining distribution. Also, we calculate the distribution locations of the correct cases as well to compare. The results are shown in Table 4. As shown in the table, it is easy to discover that the failed cases are positioned further behind the correct ones in the distribution.

C. Details of our ADR Approach

C.1. Data Rebalancing Method

The algorithm for our data rebalancing method is detailed in Algorithm 1. Initially, we calculate the sampling probability for each entity using the reverse distribution Q^r and a threshold τ. Entities with higher frequencies are assigned lower sampling probabilities, reducing the likelihood of overrepresented entities being selected. We then iterate over the entire dataset, leveraging these probabilities to filter out overrepresented instances. For each data instance d, we assess all four perspectives via random sampling. If an entity within a perspective is sampled, the perspective is marked as "pass". Instances with a number of passed perspectives greater than n_p are retained; otherwise, they are discarded.

Algorithm 1 Pseudo Code for **D**ata **R**esampling

```
# D: raw training set;
# C: target perspectives list
# tau: the threshold for entities;
# D_bal: the rebalanced data, a.k.a. D*;
# n_p, alpha: hyperparameters
D_bal=[]
for pers in C:      # build prob dict
    entity_dist =
        entity_distribution_construction
        (D,pers)
    prob_dict[pers] = {ent:tau[pers]/
        entry_dist[ent] for ent in
        entry_dict.keys()}
for instance in D: #  data rebalancing
    pass_cnt = 0
    for pers in C:
        for entity in instance['entity'
            ][pers]:
            if random.random() <
                prob_dict[pers][entity]:
                pass_cnt += 1
                break
    if pass_cnt > n_p and random.random
        () < alpha:
        D_bal.append(instance)
```

C.2. Implement Details of Data Synthesis Stage

During the Data Synthesis (DS) stage, we use ControlNet[5] [19] to generate images that closely resemble those containing tail concepts. To produce high-quality captions for the generated images, we employ ShareCaptioner[6] [1]. Finally, we leverage LLaMA 3 70B Instruct [3] to expand the captions into detailed conversations.

D. Prompts

D.1. Object Information Extraction

In this section, we release all of our prompts for guiding LLMs to do specific tasks. Firstly during the analyzing stage,

[1] stanza: link

[2] meta-llama/Meta-Llama-3-70B-Instruct: link

[3] IDEA-Research/grounding-dino-base: link

[4] Enterprise version 5.19.0: link

[5] lllyasviel/ControlNet: link

[6] Lin-Chen/ShareCaptioner: link

we utilize the LLMs to extract object information from the text within data instances at the very first step during object entity extraction. This part of the prompt we used to guide LLMs is illustrated in Figure 5.

D.2. Conversation Rewrite

We leverage LLaMA3 70B Instruct [3] to rewrite our conversations. During the Data Synthesis (DS) Stage, synthetic data and captions are generated using diffusion models and captioning models. Once the image and its corresponding caption are obtained, we employ the LM to transform the caption into a conversation. The prompt used to guide the LM is shown in Figure 6.

Moreover, during the language data synthesis process in the DS stage, we also utilize LLMs to rewrite conversations using the provided tail tokens. The corresponding prompts are shown in Figure 7. Additionally, we rewrite conversations containing tail tokens or interrogation entities (TWR in the ablation study or Section 6.2). As this task closely resembles standard rephrasing tasks with similar prompts, we will not elaborate on it further here.

References

[1] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 2

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 3

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 7

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[6] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[7] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023. 1

[8] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 7

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2023. 2

[12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1

[13] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022. 1

[14] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

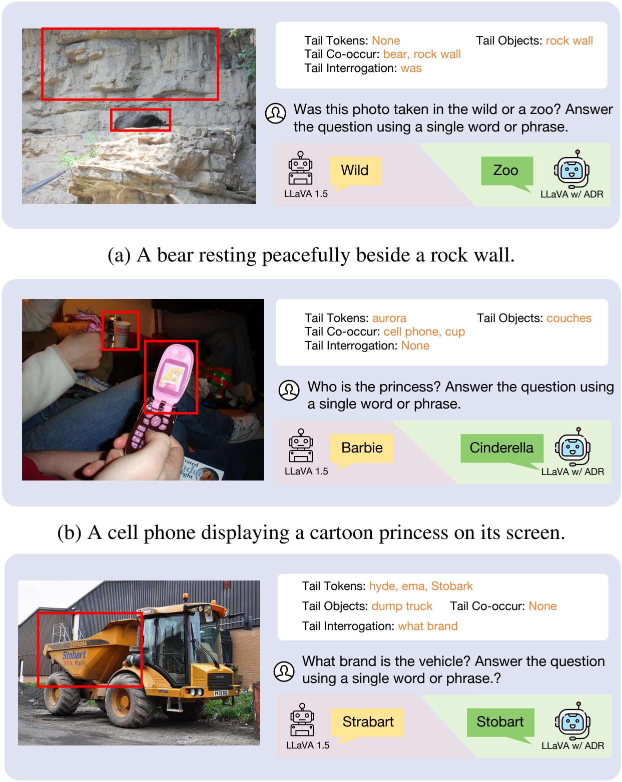
[15] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*, 2020. 2

[16] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1

[17] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. Q-bench: A benchmark for general-purpose foundation models on low-level vision. In *ICLR*, 2024. 1

[18] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 1

[19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

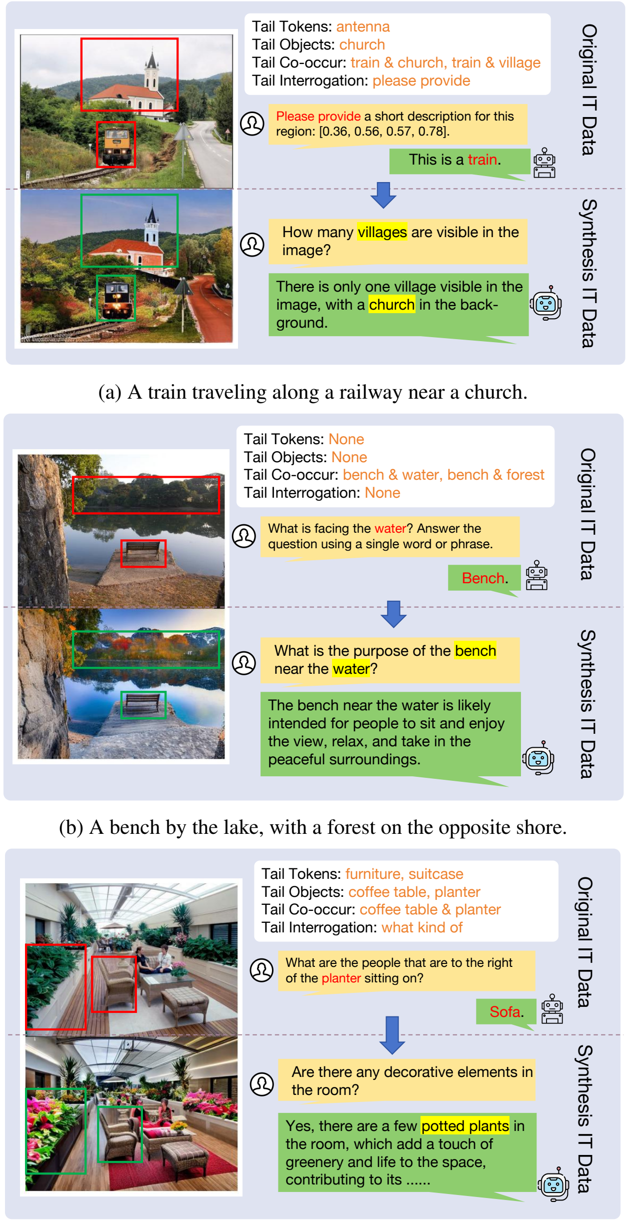

Tail Tokens: None
Tail Objects: rock wall
Tail Co-occur: bear, rock wall
Tail Interrogation: was
Was this photo taken in the wild or a zoo? Answer the question using a single word or phrase.
Wild
LLaVA 1.5
Zoo
LLaVA w/ ADR
Tail Tokens: aurora
Tail Objects: couches
Tail Co-occur: cell phone, cup
Tail Interrogation: None
Who is the princess? Answer the question using a single word or phrase.
Barbie
LLaVA 1.5
Cinderella
LLaVA w/ ADR
Tail Tokens: hyde, ema, Stobark
Tail Objects: dump truck
Tail Co-occur: None
Tail Interrogation: what brand
What brand is the vehicle? Answer the question using a single word or phrase.?
Strabart
LLaVA 1.5
Stobart
LLaVA w/ ADR

(a) A bear resting peacefully beside a rock wall.

(b) A cell phone displaying a cartoon princess on its screen.

(c) A dump truck.

Figure 1. Qualitative comparison between the baseline model (LLaVA 1.5) and our proposed method (LLaVA w/ ADR) on a few tail examples. While LLaVA 1.5 fails to answer tail questions, LLaVA w/ ADR successfully addresses them.


Tail Tokens: antenna
Tail Objects: church
Tail Co-occur: train & church, train & village
Tail Interrogation: please provide
Please provide a short description for this region: [0.36, 0.56, 0.57, 0.78].
This is a train.
Original IT Data
How many villages are visible in the image?
There is only one village visible in the image, with a church in the back-ground.
Synthesis IT Data
Tail Tokens: None
Tail Objects: None
Tail Co-occur: bench & water, bench & forest
Tail Interrogation: None
What is facing the water? Answer the question using a single word or phrase.
Bench.
Original IT Data
What is the purpose of the bench near the water?
The bench near the water is likely intended for people to sit and enjoy the view, relax, and take in the peaceful surroundings.
Synthesis IT Data
Tail Tokens: furniture, suitcase
Tail Objects: coffee table, planter
Tail Co-occur: coffee table & planter
Tail Interrogation: what kind of
What are the people that are to the right of the planter sitting on?
Sofa.
Original IT Data
Are there any decorative elements in the room?
Yes, there are a few potted plants in the room, which add a touch of greenery and life to the space, contributing to its ......
Synthesis IT Data

(a) A train traveling along a railway near a church.

(b) A bench by the lake, with a forest on the opposite shore.

(c) A furniture arrangement complemented by a variety of planters.

Figure 2. Comparison between the original instruction-tuning (IT) data and our synthesized IT data. Tail concepts in the original data are highlighted using red boxes and fonts, whereas synthesized tail concepts are marked with green boxes and yellow fonts.

Table 1. **Comparison of models trained with different approaches across multiple benchmarks.** IT represents the number of training instances used during instruction tuning. +DR signifies performance after the data rebalancing stage, and +DS indicates performance after the data synthesis stage, with the number following DS denoting the augmentation volume from the DS stage. Benchmark names are abbreviated due to space constraints. The best results are indicated in **bold**.

Method	IT*	VQA^{OK}	SEED^{2}	QB^{2}	MMS	MME^{P}	SQA^{I}	MMMU	VQA^{T}	GQA	MMB	VQA^{v2}
LLaVA 1.5	665.0K	53.2	48.7	47.3	33.5	1510.7	69.3	35.3	46.0	61.9	64.3	76.6
+DR	581.0K	55.3	57.2	46.8	33.8	1470.6	69.5	34.8	46.0	62.8	65.5	76.9
+DR +DS	665.0K	**57.4**	**57.4**	**49.6**	**35.5**	**1512.8**	**70.4**	36.7	**47.2**	**62.9**	65.0	76.9
+DS 25K	690.0K	56.2	47.5	47.9	34.5	1486.0	68.7	36.0	47.1	62.8	66.3	**77.2**
+DS 50K	715.0K	57.3	47.3	47.7	35.2	1472.5	69.9	**36.9**	47.0	62.7	**66.3**	77.1
+DS 100K	765.0K	54.5	47.2	46.1	34.6	1502.7	69.7	36.8	46.1	62.5	64.5	76.6

Table 2. Full results of ablation study on different combinations of perspectives. T, O, C, and W refer to Token, Object, Co-occurrence, and Interrogation respectively. The best results are indicated in **bold**, and the second-best results are underlined.

T	O	C	W	IT	VQA^{v2}	VQA^{T}	VQA^{OK}	GQA	SQA	SQA^{I}	REF	REF+	FLIK	POPE	SEED	Avg.
baseline				665.0K	76.6	46.0	53.2	61.9	70.4	69.3	29.4	28.5	74.9	86.9	60.6	59.8
✓				488.1K	76.5	46.6	55.3	62.3	70.8	69.2	28.5	28.1	73.8	86.7	60.2	59.8
	✓			197.9K	74.6	44.0	50.4	61.3	69.9	67.9	30.8	29.7	74.1	86.3	59.3	59.0
		✓		242.4K	75.2	43.3	47.3	61.3	70.0	68.5	31.4	29.8	76.2	86.8	59.0	59.0
			✓	176.3K	73.9	43.0	46.3	60.7	69.5	66.7	**32.3**	**31.7**	71.9	85.6	57.4	58.1
✓	✓			534.2K	76.7	47.1	55.6	62.8	71.4	68.1	30.3	29.1	75.4	86.9	60.9	60.4
✓		✓		553.4K	75.7	44.5	52.8	62.0	70.8	68.4	30.4	29.2	75.1	86.4	59.9	59.6
✓			✓	521.5K	75.7	44.5	52.8	62.0	70.8	68.4	30.4	29.2	75.1	86.4	59.9	59.6
	✓	✓		276.9K	75.4	44.6	46.8	61.7	69.0	66.4	30.6	29.4	74.2	87.1	59.3	58.6
	✓		✓	318.3K	75.7	44.6	50.9	61.8	71.5	69.0	29.9	29.0	74.9	86.8	59.6	59.4
		✓	✓	349.9K	76.8	46.8	54.4	62.5	71.5	68.8	29.9	29.2	75.7	86.8	**61.5**	60.4
	✓	✓	✓	375.9K	76.2	45.3	54.4	62.8	70.7	67.6	29.7	28.8	74.3	86.8	60.1	59.7
✓		✓	✓	575.5K	76.8	46.7	**56.7**	62.4	71.2	68.8	30.1	29.1	75.9	87.2	61.2	60.6
✓	✓		✓	559.3K	76.7	46.9	52.5	62.3	71.6	69.2	30.8	30.0	**76.6**	**87.4**	61.0	60.5
✓	✓	✓		561.5K	76.8	**47.2**	50.0	62.3	**71.7**	**69.9**	28.8	28.1	75.6	86.6	60.6	59.8
✓	✓	✓	✓	581.7K	**76.9**	46.0	55.3	**62.8**	71.4	69.5	30.2	29.7	76.2	87.2	61.0	60.6

Table 3. Full results of ablation study on different augmentation methods. Methods are introduced in Sec. 6.2. The best results are indicated in **bold**, and the second-best results are underlined.

Method	IT	VQA^{v2}	VQA^{T}	VQA^{OK}	GQA	SQA	SQA^{I}	REF	REF+	FLIK	POPE	SEED	Avg.
ALL	665.0K	76.9	**47.2**	**57.4**	62.9	**72.0**	**70.4**	30.5	29.9	76.2	86.9	61.3	61.1
Image Only	665.0K	76.9	46.5	57.2	62.5	68.8	68.4	30.6	30.2	75.9	87.3	53.8	59.8
Token Rewrite	665.0K	**76.9**	46.1	49.2	62.4	70.6	68.6	**32.3**	**31.3**	0.6	87.4	54.1	52.7
TW Rewrite	665.0K	76.9	46.9	54.9	62.5	68.9	68.7	31.0	30.3	**77.5**	87.5	53.7	59.9
PlainAug SimpAdd	665.3K	76.8	46.2	56.0	**63.0**	71.7	69.3	29.3	28.5	74.1	86.6	**61.7**	60.3
PlainAug NewCap	665.3K	76.8	46.7	54.6	62.1	68.5	69.4	31.1	30.7	77.3	**87.7**	54.1	59.9

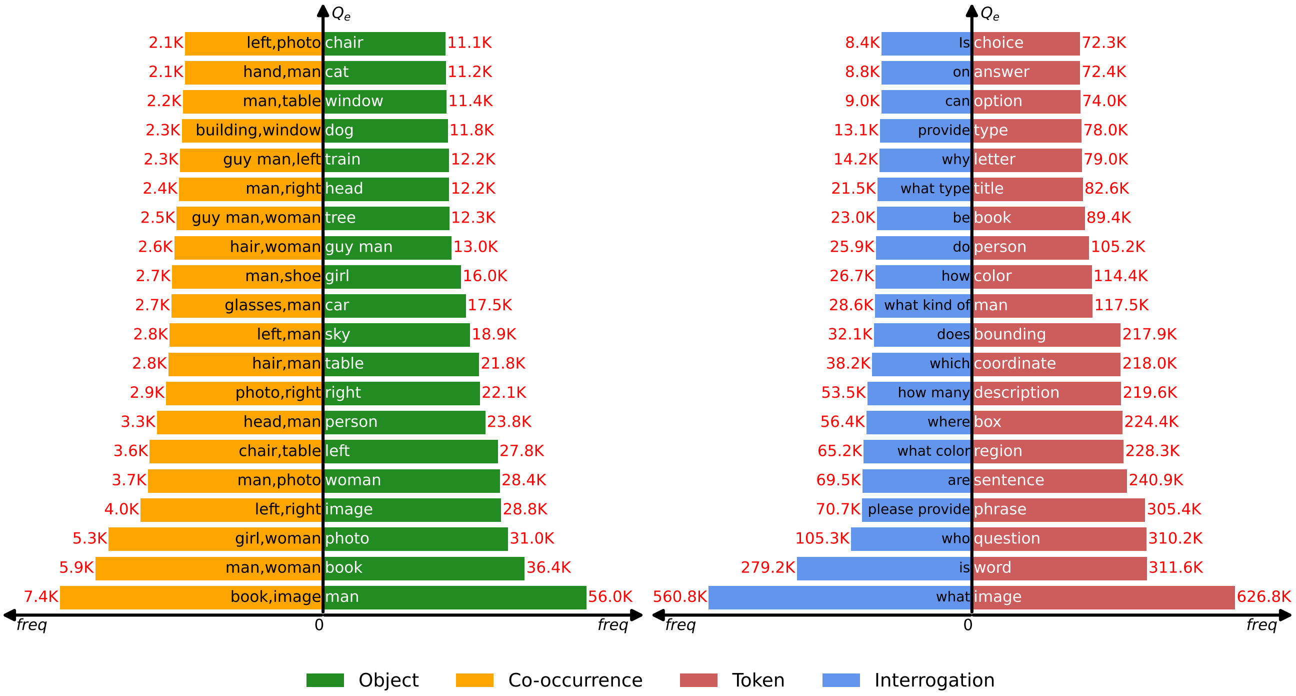
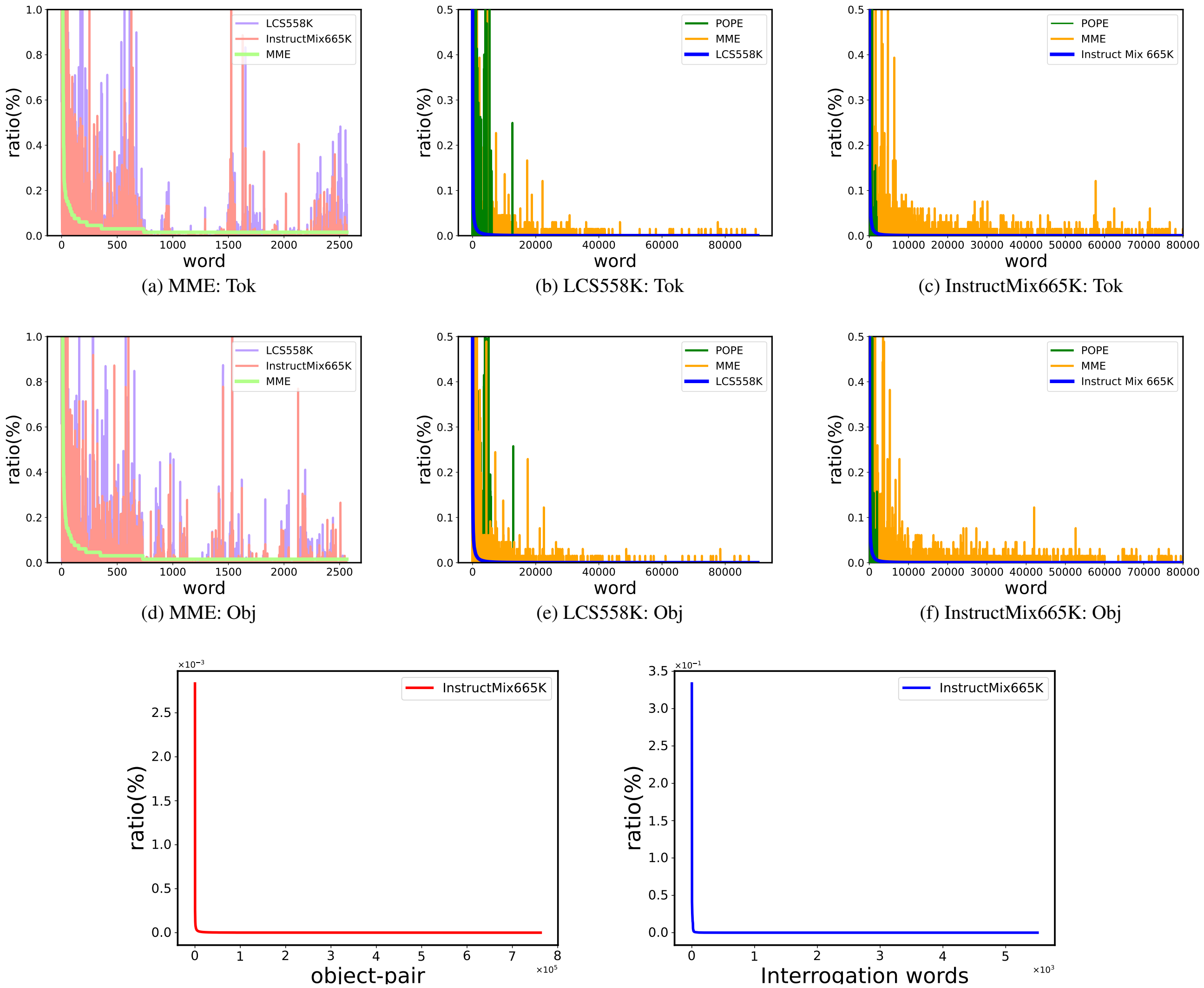


Figure 3. Top 20 most frequent entities in the instruction-tuning dataset of LLaVA 1.5.

Table 4. Distribution locations of entities in correct and incorrect answers for POPE and MME, generated by LLaVA 1.5. “Tok,” “Obj,” and “Co” refer to Token, Object, and Co-occurrence, respectively, while “W” and “C” represent wrong and correct answers, respectively. The gray rows (■) indicate the relative displacement of incorrect concepts in the distribution compared to correct concepts.

Methods	**MME**						**POPE**					
	Tok-C	**Tok-W**	**Obj-C**	**Obj-W**	**Co-C**	**Co-W**	**Tok-C**	**Tok-W**	**Obj-C**	**Obj-W**	**Co-C**	**Co-W**
Max	9738	10377	2708	3222	247315	257107	2242	2772	1085	1100	130043	141722
		+639		+514		+9792		+30		+15		+11679
Min	1	1	60	131	12732	20741	1	1	17	21	926	1033
		+0		+71		+8009		+0		+4		+107
Mean	1035	1068	842	1035	71123	79104	313	340	319	336	27457	30989
		+33		+193		+7981		+27		+17		+3532



(a) MME: Tok

(b) LCS558K: Tok

(c) InstructMix665K: Tok

(d) MME: Obj

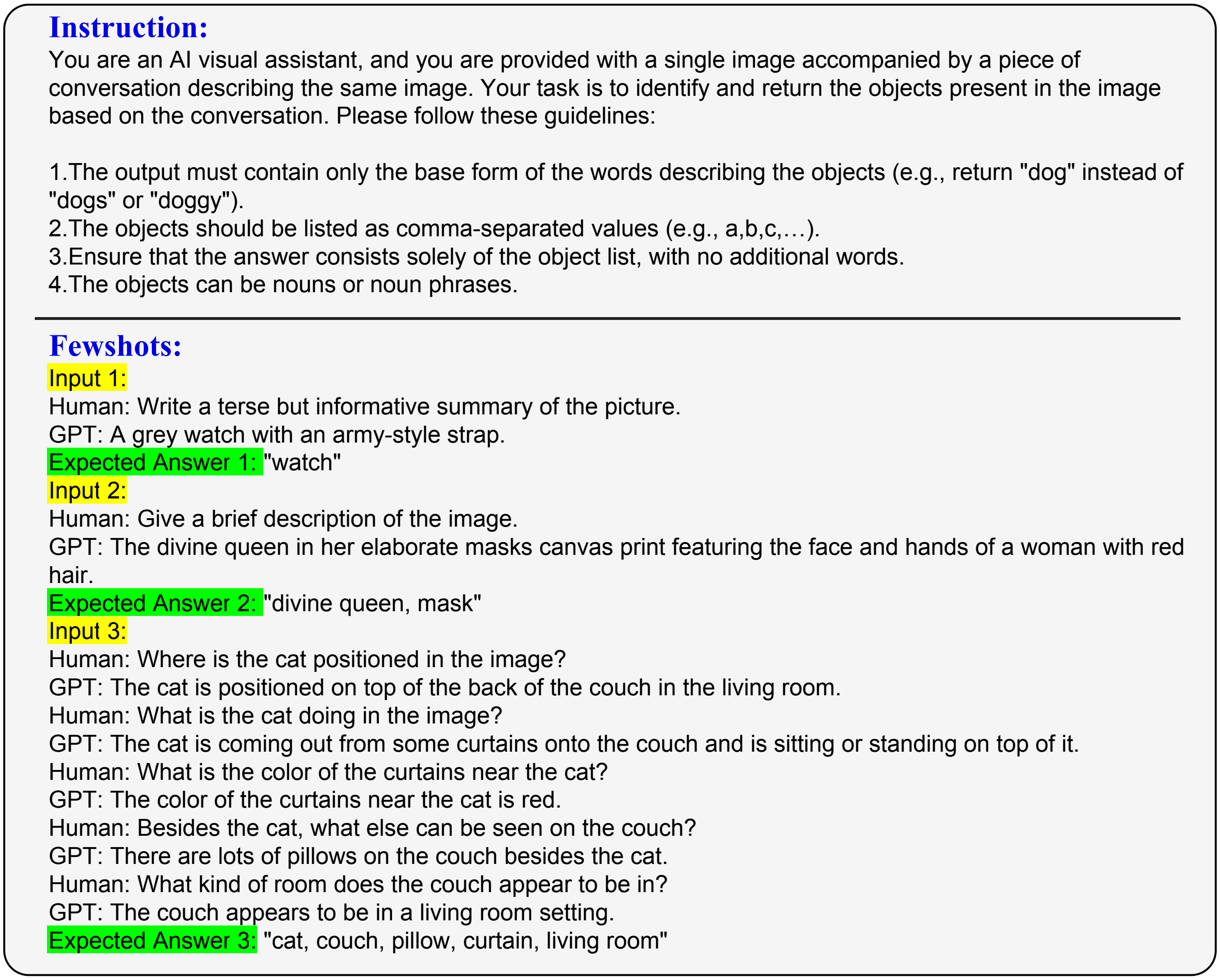
(e) LCS558K: Obj

(f) InstructMix665K: Obj

(g) InstructMix665K: Co-occurrence

(h) InstructMix665K: Interrogation

Figure 4. Long-tail distribution in instruction-tuning and benchmark datasets. Some plots feature multiple curves, with the x-axis standardized according to the dataset mentioned in the title. Distributions from various datasets are overlaid on the same graph to emphasize the differences between them. (a) Token-level word distribution in MME [4]. (b) Token-level word distribution in LCS558K [10]. (c) Token-level word distribution in InstructMix665K [10]. (d) Object-level word distribution in MME [4]. (e) Object-level word distribution in LCS558K [10]. (f) Object-level word distribution in InstructMix665K [10]. (g) Co-occurrence distribution in InstructMix665K [10]. (h) Interrogation distribution in InstructMix665K [10].



Instruction:

You are an AI visual assistant, and you are provided with a single image accompanied by a piece of conversation describing the same image. Your task is to identify and return the objects present in the image based on the conversation. Please follow these guidelines:

1.The output must contain only the base form of the words describing the objects (e.g., return "dog" instead of "dogs" or "doggy").
2.The objects should be listed as comma-separated values (e.g., a,b,c,…).
3.Ensure that the answer consists solely of the object list, with no additional words.
4.The objects can be nouns or noun phrases.

Fewshots:

Input 1:
Human: Write a terse but informative summary of the picture.
GPT: A grey watch with an army-style strap.
Expected Answer 1: "watch"
Input 2:
Human: Give a brief description of the image.
GPT: The divine queen in her elaborate masks canvas print featuring the face and hands of a woman with red hair.
Expected Answer 2: "divine queen, mask"
Input 3:
Human: Where is the cat positioned in the image?
GPT: The cat is positioned on top of the back of the couch in the living room.
Human: What is the cat doing in the image?
GPT: The cat is coming out from some curtains onto the couch and is sitting or standing on top of it.
Human: What is the color of the curtains near the cat?
GPT: The color of the curtains near the cat is red.
Human: Besides the cat, what else can be seen on the couch?
GPT: There are lots of pillows on the couch besides the cat.
Human: What kind of room does the couch appear to be in?
GPT: The couch appears to be in a living room setting.
Expected Answer 3: "cat, couch, pillow, curtain, living room"

Figure 5. Complete prompts used to guide the language model in extracting object information.

Instruction:

You are an AI visual assistant, and you are seeing a single image. What you see are provided with a few of sentences, describing the same image you are looking at. Answer all questions as you are seeing the image.

Design a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question.

Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, counting the objects, object actions, object locations, relative positions between objects, etc. Only include questions that have definite answers:

(1) one can see the content in the image that the question asks about and can answer confidently;

(2) one can determine confidently from the image that it is not in the image.

Do not ask any question that cannot be answered confidently.

Also include complex questions that are relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details.

Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

Fewshots:

Caption:

There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation:

Question: What color is the fire hydrant in the image?

Answer:The fire hydrant in the image is red.

Question:Is the fire hydrant fully covered by snow or partially visible?

Answer:The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.

Question:What is the condition of the sidewalk near the fire hydrant?

Answer:The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.

Question:What does the image suggest about the winter season in the city?

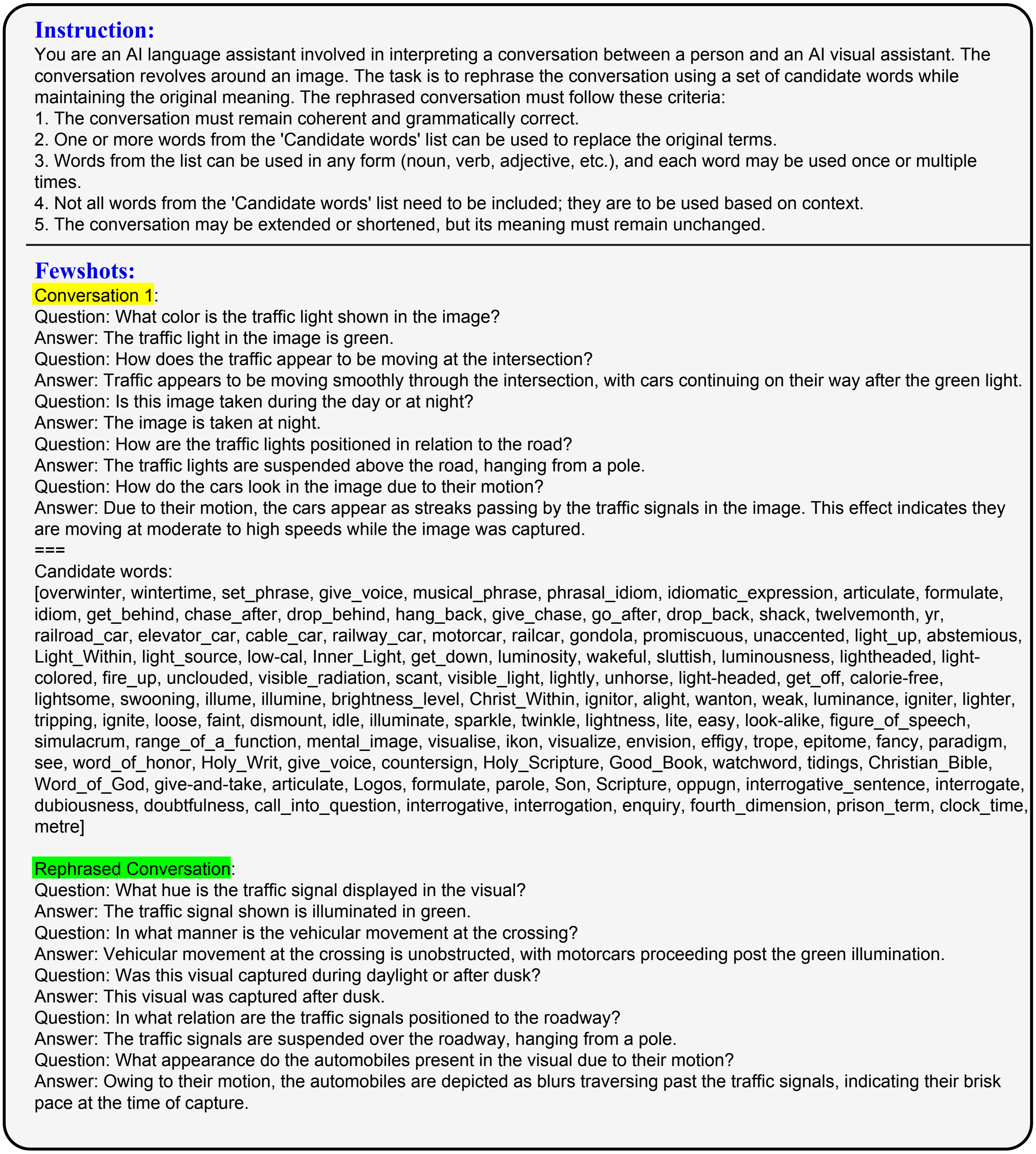
Answer:The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

Figure 6. Complete prompts used to guide the language model in converting captions into conversation instructions.



Instruction:

You are an AI language assistant involved in interpreting a conversation between a person and an AI visual assistant. The conversation revolves around an image. The task is to rephrase the conversation using a set of candidate words while maintaining the original meaning. The rephrased conversation must follow these criteria:
1. The conversation must remain coherent and grammatically correct.
2. One or more words from the 'Candidate words' list can be used to replace the original terms.
3. Words from the list can be used in any form (noun, verb, adjective, etc.), and each word may be used once or multiple times.
4. Not all words from the 'Candidate words' list need to be included; they are to be used based on context.
5. The conversation may be extended or shortened, but its meaning must remain unchanged.

Fewshots:

Conversation 1:
Question: What color is the traffic light shown in the image?
Answer: The traffic light in the image is green.
Question: How does the traffic appear to be moving at the intersection?
Answer: Traffic appears to be moving smoothly through the intersection, with cars continuing on their way after the green light.
Question: Is this image taken during the day or at night?
Answer: The image is taken at night.
Question: How are the traffic lights positioned in relation to the road?
Answer: The traffic lights are suspended above the road, hanging from a pole.
Question: How do the cars look in the image due to their motion?
Answer: Due to their motion, the cars appear as streaks passing by the traffic signals in the image. This effect indicates they are moving at moderate to high speeds while the image was captured.
===
Candidate words:
[overwinter, wintertime, set_phrase, give_voice, musical_phrase, phrasal_idiom, idiomatic_expression, articulate, formulate, idiom, get_behind, chase_after, drop_behind, hang_back, give_chase, go_after, drop_back, shack, twelvemonth, yr, railroad_car, elevator_car, cable_car, railway_car, motorcar, railcar, gondola, promiscuous, unaccented, light_up, abstemious, Light_Within, light_source, low-cal, Inner_Light, get_down, luminosity, wakeful, sluttish, luminousness, lightheaded, light-colored, fire_up, unclouded, visible_radiation, scant, visible_light, lightly, unhorse, light-headed, get_off, calorie-free, lightsome, swooning, illume, illumine, brightness_level, Christ_Within, ignitor, alight, wanton, weak, luminance, igniter, lighter, tripping, ignite, loose, faint, dismount, idle, illuminate, sparkle, twinkle, lightness, lite, easy, look-alike, figure_of_speech, simulacrum, range_of_a_function, mental_image, visualise, ikon, visualize, envision, effigy, trope, epitome, fancy, paradigm, see, word_of_honor, Holy_Writ, give_voice, countersign, Holy_Scripture, Good_Book, watchword, tidings, Christian_Bible, Word_of_God, give-and-take, articulate, Logos, formulate, parole, Son, Scripture, oppugn, interrogative_sentence, interrogate, dubiousness, doubtfulness, call_into_question, interrogative, interrogation, enquiry, fourth_dimension, prison_term, clock_time, metre]

Rephrased Conversation:
Question: What hue is the traffic signal displayed in the visual?
Answer: The traffic signal shown is illuminated in green.
Question: In what manner is the vehicular movement at the crossing?
Answer: Vehicular movement at the crossing is unobstructed, with motorcars proceeding post the green illumination.
Question: Was this visual captured during daylight or after dusk?
Answer: This visual was captured after dusk.
Question: In what relation are the traffic signals positioned to the roadway?
Answer: The traffic signals are suspended over the roadway, hanging from a pole.
Question: What appearance do the automobiles present in the visual due to their motion?
Answer: Owing to their motion, the automobiles are depicted as blurs traversing past the traffic signals, indicating their brisk pace at the time of capture.

Figure 7. Complete prompts used to guide the language model in rewrite conversation instructions using given tokens.