Unveiling the Ignorance of MLLMs: Seeing Clearly, Answering Incorrectly -Supplementary Material-

Abstract

Due to the space constraints of the main manuscript, this Supplementary Material provides a comprehensive presentation of additional details and analyses. The content is structured as follows: (1) a detailed overview of the MMVU test set, including its construction criteria and a preliminary evaluation conducted on a subset of the test set; (2) findings from initial exploratory experiments to assess the test set's effectiveness using a subset of the MMVU dataset; (3) an examination of the MLLM's performance under various data combination strategies, prompt versions, and subsets of MMVU-Train of differing sizes, accompanied by an in-depth discussion on the construction of training prompts and their impact on model performance; (4) the training of an MLLM using the MMVU-Train dataset, validated on both a subset of the MMVU test set and other general datasets; (5) representative examples of paired positive and negative samples; and (6) the specific content of the "Instruction" prompt employed during model testing.

Contents

1. MMVU Test set Details

It is important to emphasize that the MMVU-Train dataset is automatically constructed, while the MMVU test set is manually designed.

1.1. Category Definition and Negative Question Design

Character-Level

1. Character/Number: Recognize characters and numbers in an image, and some reasoning is based simply on the content of the characters or numbers in the image.

Adding, deleting, modifying, or swapping positions of characters or numeric content in an image.

Attribute-Level

1. Presence: Determine whether a person, animal, or object is present in the image.

Ask questions by replacing or deleting people, animals, or objects in the image.

2. Color/Texture: Determine whether the color or texture of an object in an image is correct.

Replace the color of an object in an image or replace it with an object of similar texture to ask questions.

3. Number: Count the number of people, animals, or objects in an image.

Modify the number of people, animals, or objects in the image to ask questions.

4. Shape: Determine if an object is shaped correctly.

Modify the shape of objects in the image to ask questions.

5. Posture: Determine if a person, animal, or object is in the correct posture.

Ask questions by substituting gestures of characters, animals, or objects.

6. Position: Determine whether the absolute or relative position of a person, animal, or object is correct.

Change the absolute or relative position of a character, animal, or object to ask a question.

Context-Level

1. Abstract Knowledge: Assess understanding of abstract concepts such as emotions, aesthetics, connotations, symbols, culture, news, allusions, legends, common sense, functions, etc.

No abstract knowledge is involved, and questions are asked only visually through images.

2. Concrete Knowledge: Evaluate specific factual information or well-defined objects and scenarios depicted in the image, such as landmarks, celebrities, well-known objects, etc.

Changing the attributes or names of landmarks, famous objects, or famous people.

3. Expertise: Test specialized knowledge related to specific fields or domains illustrated in the image, such as specific knowledge in various vertical fields, industries, disciplines, etc.

Ask questions with the wrong expertise.

4. Activity: Identify and interpret actions or events occurring within the image, requiring a dynamic understanding of the scene.

Ask questions with the wrong activity.

5. Relationships: Analyze and comprehend the interactions or relationships between multiple entities within the image, such as social dynamics or physical connections.ch as social dynamics or physical connections.

Change the relationship between entities in an image to ask questions.

1.2. Preliminary Evaluation Details.

We evaluated several top-performing multimodal large language models (MLLMs) on the MMVU test set, using the official inference code and parameter settings for each model to ensure optimal performance during the inference process, without any adjustments to the parameter settings.

Figure 1 illustrates that the model's accuracy in responding to negative questions is lower than its performance on positive questions. When evaluating only positive questions, as is common in most existing benchmarks, many models achieve an accuracy rate exceeding 70%, indicating strong performance. However, the model's overall performance on negative questions remains suboptimal, suggesting that current models continue to face challenges with leading questions. The higher accuracy on positive questions implies that the models can effectively process and understand the content of images. In contrast, the lower accuracy on negative questions indicates a tendency to overlook visual content when confronted with leading questions, resulting in incorrect responses. This underscores the importance of the MMVU test set in quantitatively assessing the robustness of models against leading questions.

2. Discussion about the Effectiveness of the MMVU Test Set

Is the proposed MMVU test set effective for MLLMs? MMStar [5] critically examines prevailing evaluation methods and brings to light a significant issue: many instances don't require visual stimuli for accurate responses. Answers can be inferred directly from the questions, choices, or the intrinsic world knowledge within LLMs. This trend pervades existing benchmarks. To investigate if this issue persists in our benchmark, we conduct experiments using both image and text inputs, as well as only text inputs. The results, showcased in Tab. 1 and Tab. 2, reveal that without text, the model's performance on our benchmark dips below random selection. This underscores the necessity of multimodal input for our proposed benchmark and highlights that our benchmark is not solely reasoned out by text.

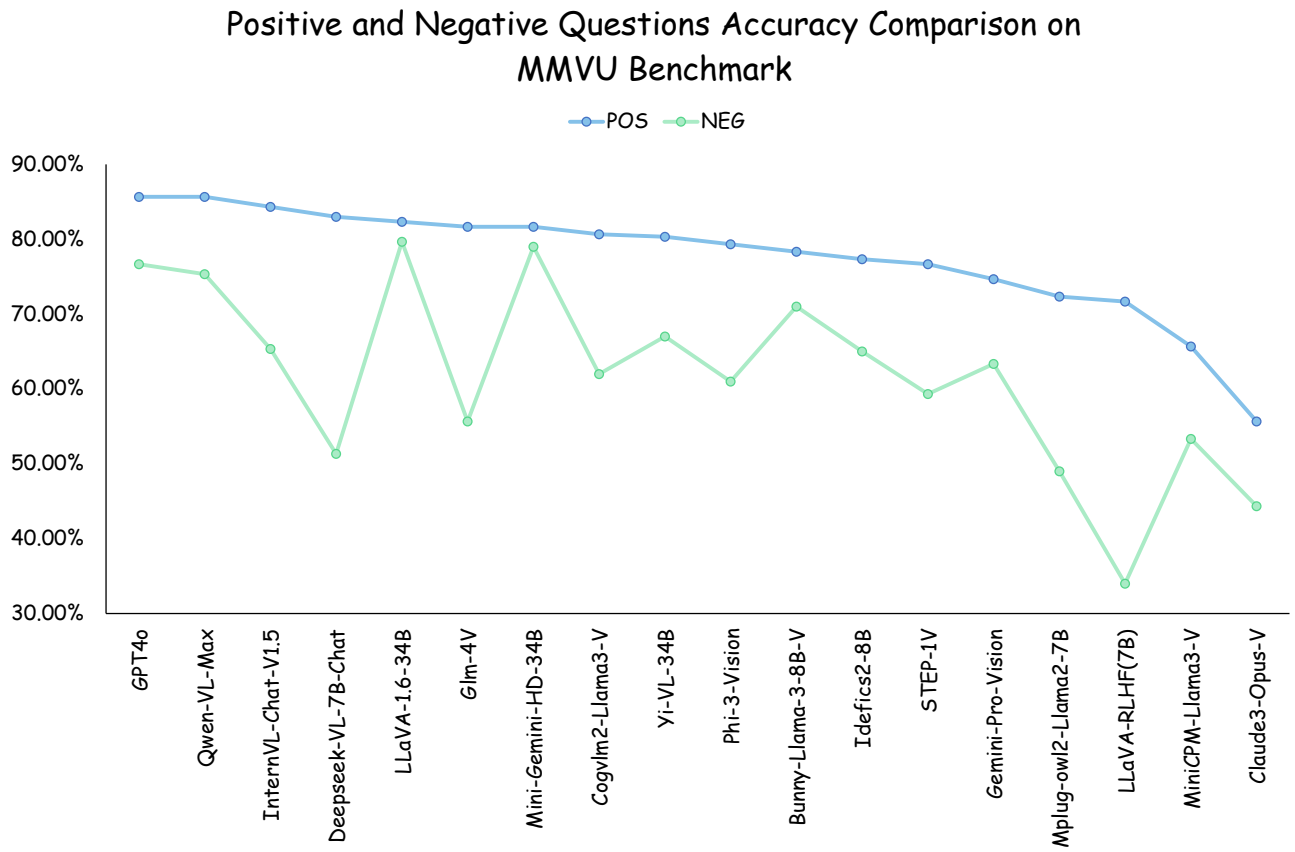


Figure 1. Visualization of fine-tuning loss of different data composition (version 0).

Table 1. Performance of MLLMs with images as input on the MMVU test set. §We evaluate the officially released checkpoint by ourselves. Abbreviations: Char/Num. (Character/Number), Pres. (Presence), Color/Tex. (Color/Texture), Num. (Number), Shape (Shape), Posture (Posture), Pos. (Position), Abstract. (Abstract Knowledge), Concrete. (Concrete Knowledge), Expert (Expertise), Act. (Activity), Rel. (Relationships). ↑: higher is better, ↓: lower is better.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA↑
Random	25	25	25	25	25	25	25	25	25	25	25	25	25
Phi-3-vision (4B) [1]	62.50	59.09	58.33	37.50	70.83	33.33	31.82	54.55	66.67	41.94	58.33	50.00	52.33
Bunny-Llama-3-8B-V [13]	55.00	63.64	54.17	37.50	79.17	62.50	54.55	72.73	85.71	48.39	75.00	50.00	60.67
Idefics2-8B [19]	57.50	59.09	54.17	50.00	79.17	41.67	27.27	77.27	76.19	45.16	75.00	40.91	56.67

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Random	50	50	50	50	50	50	50	50	50	50	50	50	50
Phi-3-vision (4B) [1]	19.35	18.75	26.32	43.75	19.05	55.56	58.82	40.00	22.22	48.00	33.33	31.25	34.03
Bunny-Llama-3-8B-V [13]	15.38	22.22	18.75	40.00	5.00	28.57	29.41	23.81	10.00	40.00	10.00	26.67	22.22
Idefics2-8B [19]	23.33	27.78	23.53	20.00	13.64	50.00	40.00	22.73	11.11	41.67	14.29	40.00	26.72

3. MMVU-Train Data Construction Preliminary Experiment

3.1. Notes

1. *It is important to note that, to reduce the validation period, the test set used in the preparatory experiments does not*

Table 2. Performance of MLLMs without images as input on the MMVU test set. §We evaluate the officially released checkpoint by ourselves. Abbreviations: Char/Num. (Character/Number), Pres. (Presence), Color/Tex. (Color/Texture), Num. (Number), Shape (Shape), Posture (Posture), Pos. (Position), Abstract. (Abstract Knowledge), Concrete. (Concrete Knowledge), Expert (Expertise), Act. (Activity), Rel. (Relationships). ↑: higher is better, ↓: lower is better.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA↑
Random	25	25	25	25	25	25	25	25	25	25	25	25	25
Phi-3-vision (4B) [1]	15.0	13.64	20.83	0.0	12.5	8.33	9.09	27.27	19.05	16.13	12.5	13.64	14.00
Bunny-Llama-3-8B-V [13]	22.5	13.64	16.67	4.17	4.17	16.67	22.73	22.73	28.57	19.35	8.33	27.27	17.33
Idefics2-8B [19]	15.0	4.55	8.33	4.17	8.33	4.17	0.0	31.82	23.81	3.23	0.0	13.64	9.76

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Posture	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Random	50	50	50	50	50	50	50	50	50	50	50	50	50
Phi-3-vision (4B) [1]	72.73	72.73	58.33	100.0	66.67	75.0	81.82	57.14	60.0	72.22	62.5	72.73	69.34
Bunny-Llama-3-8B-V [13]	64.0	75.0	63.64	83.33	91.67	69.23	44.44	64.29	50.0	70.0	80.0	57.14	67.09
Idefics2-8B [19]	77.78	90.91	80.0	85.71	85.71	92.31	100.0	56.25	61.54	95.24	100.0	80.0	82.10

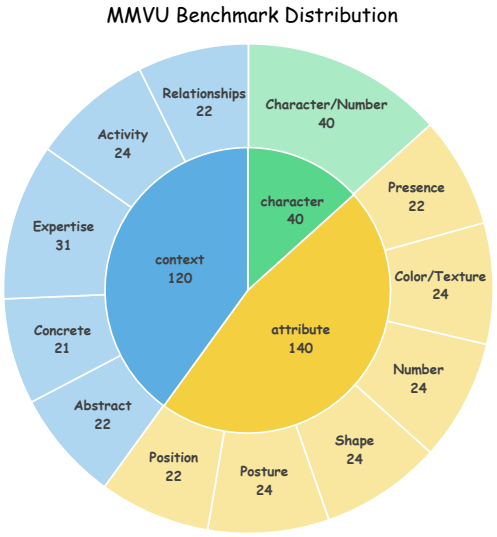


Figure 2. Distribution of MMVU benchmark.

constitute the full MMVU test set. The distribution of this test set, depicted in Fig. 2, closely resembles that of the final test set, with the primary difference being the scaling of each category. Consequently, the phenomena observed in the preparatory experiments on this subset are consistent with those observed using the full MMVU test set.

2. *It should be noted that our data construction process is not carried out in multiple steps; rather, it is implicitly performed in a single step.*

3.2. Preliminary Testing Data Distribution.

We have collected images that matched 12 subcategory themes from the web and some test sets, such as Flickr-30k [35], and facilitated setting misleading prompts. We meticulously design questions and standard answers based solely on the image information without relying on the original captions. The distribution of the subset of the MMVU test set is shown in Fig. 2. It categorizes questions into three main types: ”context” (120 questions), “character” (40 questions), and “attribute” (140 questions). The “context” category includes subtypes such as relationships (22), activity (24), expertise (31), concrete (21), and abstract (22) questions. The “character” category comprises character/number questions (40). The “attribute” category encompasses presence (22), color/texture (24), number (24), shape (24), posture (24), and position (22) questions. This distribution highlights a comprehensive and balanced approach, ensuring a well-rounded benchmark with detailed subcomponents across various dimensions of question types.

3.3. Training Details

We employed a two-stage training strategy with specific hyperparameters for pre-training and fine-tuning, as detailed in Tab. 3. For testing, we used greedy search to ensure reproducibility.

3.4. Discussion 1: Can We Simply Introduce Negative Samples to Train MLLMs?

We observed that many MLLMs could understand the content but struggled to answer questions. We hypothesized that the insufficiency of negative samples might be the cause. To address this, we generated additional negative samples and

Table 3. Hyperparameters.

	batch size	lr	lr schedule	lr warmup ratio	weight decay	epoch	optimizer	DeepSpeed stage
Pretrain	256	1e-5	cosine decay	0.03	0	1	AdamW	2
Finetune	128	2e-4	cosine decay	0.03	0	1	AdamW	3

incorporated them into the training process. We sampled 7k and 24k samples from LLaVA and used GPT-4V to generate the negative samples.

Prompt. We employ two types of prompts for this task *version 0*, which restricts only the number of problems, and *version 1*, which restricts both the type and number of problems. Their common prompt is partially displayed in Fig. 17, while their respective different prompts are displayed in Fig. 4 and Fig. 5, respectively.

3.4.1. Training and Results of the *Version 0*.

We train MLLMs using the sampled data and synthetically generated data. The hyper-parameters used for training remain consistent with those outlined in Tab. 3. We first explore the model's performance using different data ratios: 1) sampled data alone, 2) sampled data concatenated with generated data, and 3) sampled data combination with generated data. The loss curves for the fine-tuning process are illustrated in Fig. 6. The loss curves in Fig. 6 reveal that incorporating generated negative data, whether through concatenation or combination, leads to a further reduction in model loss compared to training solely on sampled data. Notably, the lowest loss is achieved when utilizing a combination of both sampled and generated data.

Tab. 4 and Tab. 5 present the model's performance on the MMVU test set and other benchmarks across the different data compositions. As expected, training exclusively on sampled data yields the lowest overall performance. Interestingly, while splicing sampled data with generated negative samples leads to a performance decrease on the generic benchmark, it demonstrates improvements on certain phantom benchmarks. The most effective approach proved to be combining the original sampled data with generated negative samples. This strategy consistently enhances performance across most benchmarks, often by a significant margin.

Ablation about conversation rounds. Our initial concatenation approach directly appended all conversation rounds of a negative sample to the original data. However, this might lead to performance degradation due to data imbalance. To investigate this, we conducted an ablation study on the 24k dataset, focusing on the number of conversation rounds included from negative samples. Fig. 7 presents the model's training loss across different numbers of conversation rounds. Additionally, Tab. 7 and Tab. 6 showcase the corresponding performance on the MMVU and generalized benchmarks, respectively. The experimental results indicate that using fewer concatenated conversation rounds does not significantly degrade model performance. However, a minor performance drop is still observed. We hypothesize that this drop stems from the increased variability within each training sample when fewer rounds are used. This variability might make it more challenging for the model to learn consistent patterns and generalize effectively.

3.4.2. Training and Results of the *Version 1*.

To further investigate the model's performance under constraints on question types, we introduce a modified prompt and generate a new 24k dataset for experiments. Interestingly, despite prompting for negative samples, GPT-4V still produced a mix of positive and negative examples, with an approximate ratio of 1:13. Utilizing only this generated dataset, we have conducted experiments to analyze the impact of different problem types. Fig. 8 illustrates the training loss, while Tab. 8 and Tab. 9 present the corresponding performance results on benchmarks.

This experiment yielded several key insights: **1) Avoid rigidly specifying the number of question types, particularly for characteristics like characters.** Enforcing such constraints, especially when not all images contain those characteristics, can lead to generating nonsensical question-answer pairs and consequently, performance degradation. **2) Training solely on generated data can be surprisingly effective, but only when both positive and negative sample pairs are present.** This finding highlights the potential of synthetic data for training MLLMs. Notably, using only negative samples restricts the model from producing negative answers, emphasizing the importance of balanced datasets. Consequently, we shifted our focus to **generating paired positive and negative samples** to investigate model performance under this training paradigm further.

3.5. Discussion 2: What Constitutes an Effective Pair of Positive and Negative Samples for Training?

Having established the benefit of incorporating both positive and negative samples during training, we now delve into effective construction methods for such pairs. We investi-

STEP-1: Parse and extract information based on the Input JSON and Information Topics:

1. Determine if there is any very clear text or numbers (You must make sure the identification is right) in the image. If yes, extract the corresponding text or numbers along with their position.
2. Determine if there is any object in the image, including the properties of that object, such as name/category/color/texture/shape/pose (The properties of an object may not be unique, such as color, it may be a mixture of different colors), position, and quantity (You must make sure the identification is right).
3. Determine if there is any person in the image, including attributes such as name, position, gender, and facial expression (You must make sure the identification is right. You must differentiate between a portrait in a photo or poster and a real person, don't confuse them.).
4. Understand the events occurring in the image at a global level, and assess whether they relate to culture, emotions, or knowledge (You must make sure the identification is right).
5. Understand the relationships between objects within the image, such as relative positions or hierarchical relationships (You must make sure the identification is right).
6. Understand the relationships between people and objects, such as someone riding a bike or a person's actions (You must make sure the identification is right).
7. If there are two or more people present, extract activities they are engaged in.

Figure 3. Common Prompt for information extraction.

Table 4. Ablation about only introducing negative samples on the MMVU test set. Concat. and Neg. denotes concatenate and negative samples, respectively.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA↑
LLaVA 7K	2.50	0.00	4.17	0.00	0.00	0.00	4.55	0.00	4.76	0.00	0.00	4.55	1.67
LLaVA 7K Concat. Neg.	2.50	0.00	4.17	4.17	20.83	0.00	4.55	0.00	0.00	3.23	4.17	4.55	4.00
LLaVA 7K Conbine Neg.	5.00	4.55	8.33	16.67	25.00	4.17	9.09	18.18	4.76	3.23	16.67	9.09	10.00
LLaVA 24K	17.50	22.73	20.83	12.50	45.83	16.67	13.64	50.00	28.57	29.03	45.83	31.82	27.33
LLaVA 24K Concat. Neg.	20.00	13.64	8.33	16.67	58.33	12.50	4.55	40.91	19.05	9.68	25.00	31.82	21.33
LLaVA 24K Conbine Neg.	27.50	27.27	25.00	20.83	54.17	16.67	9.09	45.45	38.10	25.81	54.17	36.36	31.33
Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
LLaVA 7K	83.33	100.00	80.00	100.00	100.00	100.00	75.00	100.00	80.00	100.00	100.00	80.00	93.06
LLaVA 7K Concat. Neg.	88.89	100.00	66.67	87.50	44.44	100.00	66.67	100.00	100.00	92.31	75.00	75.00	84.42
LLaVA 7K Conbine Neg.	86.67	83.33	71.43	66.67	57.14	92.31	66.67	66.67	92.31	93.75	55.56	81.82	77.61
LLaVA 24K	72.00	54.55	68.75	75.00	38.89	77.78	62.50	45.00	68.42	62.50	38.89	46.15	59.41
LLaVA 24K Concat. Neg.	52.94	62.50	85.71	63.64	26.32	80.00	87.50	50.00	77.78	83.33	53.85	50.00	63.01
LLaVA 24K Conbine Neg.	59.26	50.00	60.00	58.33	31.58	78.95	77.78	47.37	60.00	68.00	35.00	42.86	55.45

gate two distinct approaches: **1) Explicit Positive/negative Prompting.** This method involves directly prompting the model to generate both positive and negative samples, with answers explicitly labeled as "yes" or "no." This approach offers straightforward control over the positive/negative balance in the training data (*version 2*). **Confounding Answer Construction.** This approach focuses on crafting questions that naturally lend themselves to multiple plausible answers, without explicitly forcing a "yes/no" response. This strategy aims to create a more nuanced and challenging training dataset, potentially enhancing the model's ability to discern subtle differences in meaning and context (*version 3*). The GPT-4V prompts for version 2 and version 3 are shown in Fig. 9 and Fig. 10, respectively. To evaluate the effectiveness of these two versions of prompts, we generated 24k positive-negative sample pairs using each version's prompt. We then incorporated these datasets into our training pipeline, maintaining consistent experimental settings with previous trials. Fig. 11 presents the training loss curves observed for both approaches, while Tab. 10 and Tab. 11 detail the corresponding

STEP-2: Generate questions and corresponding correct answers based on the extracted information in STEP-1, Input Image, and Question Types (Do not output):
Question Types:

1. If there is no text/number recognized or you do not make sure the text/number is correct, do not ask this type of question. If text is present, inquire about modifications to characters or numbers: Generate questions involving minor alterations, such as reversing the order of digits (modifying a number) or replacing a character.
2. If characters or numbers are present in the image, modify one letter of a word or one digit of a number, then ask if the modified character or number exists, and add a distractor sentence to make the question more challenging. Otherwise, do not ask this type of question.
3. Ask if certain objects exist: Replace the nouns of objects in the image with visually similar but nonexistent objects, then ask if the model identifies the object.
4. Ask if objects exist and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to make the question more challenging.
5. If objects have attributes such as names of people, colors of objects, counts of objects, or positions, ask modified questions about object attributes and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to guide the model to make incorrect inferences or answers.
6. Ask global understanding questions: Understand the broader context, culture, emotions, and knowledge depicted in the image. Pose questions that differ slightly from true understanding but are very similar.
7. Ask global understanding questions and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to make the question more challenging.
8. Ask local understanding questions: Based on the extracted relationships between objects, between people and objects, or between people, actions, or relative positions, pose local understanding challenges (negative samples).
9. Ask local understanding questions and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to guide the model to make incorrect inferences or answers.

Requirements:

1. Generate 7-10 questions.
2. Ask directly, without adding any information such as "in the conversation," "mentioned in the conversation," or "in the image."
3. Answer directly, with explanation, succinctly, with responses not exceeding 30 characters.
4. Questions must be real and answerable. Answers must be correct, definitely,and not speculative.
5. Do not generate questions that the answer is "uncertain".
6. The format of the generated questions and answers is: 'conversations': ['from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer', 'from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer', ..., 'from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer']

STEP-3: Output JSON sample

Figure 4. GPT4 prompt for *version 0*.

performance results across various benchmarks. Our experimental results demonstrate that utilizing prompt version 3 to generate paired positive and negative samples yielded the strongest overall performance. Consequently, we adopted version 3 as our preferred method for generating training data in the main experiments.

4. Training an MLLM on the Proposed MMVU-Train

4.1. Framework.

The structure consists of a visual encoder, a visual-language connector, and a language model. Specifically, we employ SigLIP [37] as the visual encoder, Phi-2 (2.7B) [29] as the language model, and a two-layer MLP as the connector. We build the model based on Bunny [13].

4.2. Training Scheme.

Given an image and text, the image is processed through a visual encoder to obtain visual features. These features are then adjusted via a connector to align with the dimensions of the language model. The text is tokenized to generate textual features. These visual and textual features are concatenated and fed into the language model for generating responses. During training, each sample comprises instruction and response. The instructions are masked, and only the response and the model's output are used to calculate the loss.

We employ the two-stage training strategy. In the first

STEP-2: Generate questions and corresponding correct answers based on the extracted information in STEP-1, Input Image, and Question Types (Do not output):
Question Types:

1. **Character:**
   (a) If no text/number is recognized or the text/number recognized is not sure, do not ask questions about text/number. If text exists, inquire about modifications to characters or numbers: Generate questions involving minor alterations, such as reversing the order of digits (modifying a number) or replacing a character.
   (b) If there are characters or numbers in the image, modify one letter or one digit, then ask if the modified character or number exists, and add a distractor sentence to make the question more challenging. Otherwise, do not ask this type of question.
2. **Semantic:**
   (a) Ask if certain objects exist: Replace the nouns of objects in the image with visually similar but nonexistent objects, then ask if the model identifies the object.
   (b) Ask if objects exist and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to make the question more challenging.
   (c) If objects have attributes such as names of people, colors of objects, counts of objects, or positions, ask modified questions about object attributes and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to guide the model to make incorrect inferences or answers.
3. **Understanding:**
   (a) Ask global understanding questions: Understand the broader context, culture, emotions, and knowledge depicted in the image. Pose questions that differ slightly from true understanding but are very similar.
   (b) Ask global understanding questions and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to make the question more challenging.
   (c) Ask local understanding questions: Based on the extracted relationships between objects, between people and objects, or between people, actions, or relative positions, pose local understanding challenges (negative samples).
   (d) Ask local understanding questions and add a distractor sentence: Add a distractor sentence to questions based on modifying characters or numbers to guide the model to make incorrect inferences or answers.

Requirements:

1. Generate 2 character-type questions, 4 semantic-type questions, and 2 understanding-type questions.
2. Ask directly, without adding any information such as "in the conversation," "mentioned in the conversation," or "in the image."
3. Answer directly, with explanation, succinctly, with responses not exceeding 30 characters.
4. Questions must be real and answerable. Answers must be correct, definitely,and not speculative.
5. Do not generate questions that the answer is "uncertain".
6. The format of the generated questions and answers is: 'conversations': ['from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer', 'from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer', ..., 'from': 'human', 'value': 'Question', 'from': 'gpt', 'value': 'Answer']

STEP-3: Output JSON sample

Figure 5. GPT4 Prompt for *version 1*.

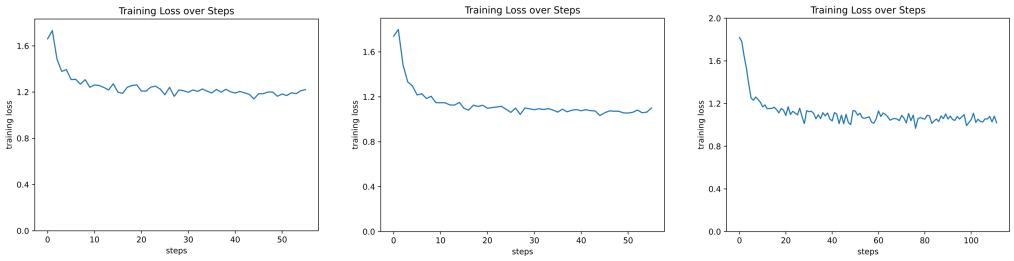
stage, the visual and language alignment stage, we use *Bunny-pretrain-LAION-2M* [13] data to train only the connector while freezing the vision encoder and the language model. In the second stage, we fine-tune both the connector and the language model using the generated MMVU dataset (MMVU-Train). To further enrich the dataset's diversity, we include OCR data (OCR-VQA [30]), visual question-answering data (VQA-v2 [12], GQA [15], OK-VQA [27], A-OKVQA [31], Visual Genome [18], and RefCOCO [17]). Additionally, to maintain the language capabilities of the MLLM, we incorporated text-only data from WizardLM [33].

4.3. Implementation Details.

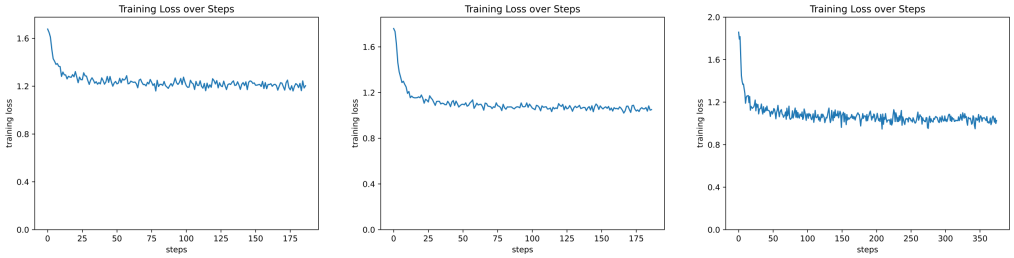
In the visual and language alignment stage, we train only the connector for one epoch with a learning rate of 1e-5. In the fine-tuning phase, we fine-tune both the connector and the language model using LoRA [14] with a learning rate of 2e-4. The training is implemented in PyTorch using 8 Nvidia A100 GPUs in an internal server.

4.4. Evaluation Benchmark.

We evaluate the performance of different MLLMs with the MMVU test set (MMVU^{A} and MMVU^{M} denote the average score of RA and MR, respectively.), and general benchmarks.



LLaVA 7K | LLaVA 7K Concat Neg | LLaVA 7K Conbine Neg



LLaVA 24K | LLaVA 24K Concat Neg | LLaVA 24K Conbine Neg

Figure 6. Visualization of fine-tuning loss of different data composition (version 0).

Table 5. Ablation about only introducing negative samples on general benchmarks. Concat. and Neg. denotes concatenate and negative samples, respectively.

Dataset	MME^P	MME^C	MMB^T	MMB^D	POPE
LLaVA 7K	719.07	272.86	1.68	2.23	74.70
LLaVA 7K Concat. Neg.	1095.19	267.5	0.50	1.03	82.69
LLaVA 7K Conbine Neg.	883.14	287.14	4.76	4.90	82.99
LLaVA 24K	1039.62	265.36	41.48	42.53	82.51
LLaVA 24K Concat. Neg.	1308.59	254.64	15.53	18.21	82.22
LLaVA 24K Conbine Neg.	1174.62	270.35	49.61	53.78	84.70

We utilize commonly used general benchmarks: MME perception (MME^P) [10], MME cognition (MME^C) [10], MMBench test split (MMB^T) [26], MMBench dev split (MMB^D) [26], SEED-Bench-1 (SEED) [20], MMMU validation split ($MMMU^V$) [36], MMMU test split ($MMMU^T$) [36], VQA-v2 test-dev split [12], GQA test-dev-balanced split [15], and the average F1-score across random, popular, and adversarial categories on the validation set of MSCOCO (POPE) [22]. This comprehensive validation ensures robust evaluation across diverse metrics and scenarios.

4.5. Expriment Results

Comparison of MLLMs. We augment our generated MMVU-Train with additional datasets to form a comprehensive new dataset. We validate this expanded dataset on the proposed MMVU and several general benchmarks. As shown in Tab. 12, Experiments show that the 3-billion parameter model trained on this enriched dataset outperforms the LLaVA-v1.5-7B on most benchmarks. Remarkably, it also exceeds the performance of some models with 13-billion parameters on several benchmarks.

Ablation about combining the proposed MMVU-Train with other datasets. We conduct experiments using an expanded dataset to evaluate the impact of integrating the MMVU-Train with other datasets. We sample 24k instances from the LLaVA 158k dataset and combine them with other datasets to construct 106k samples (Mix 106k), maintaining

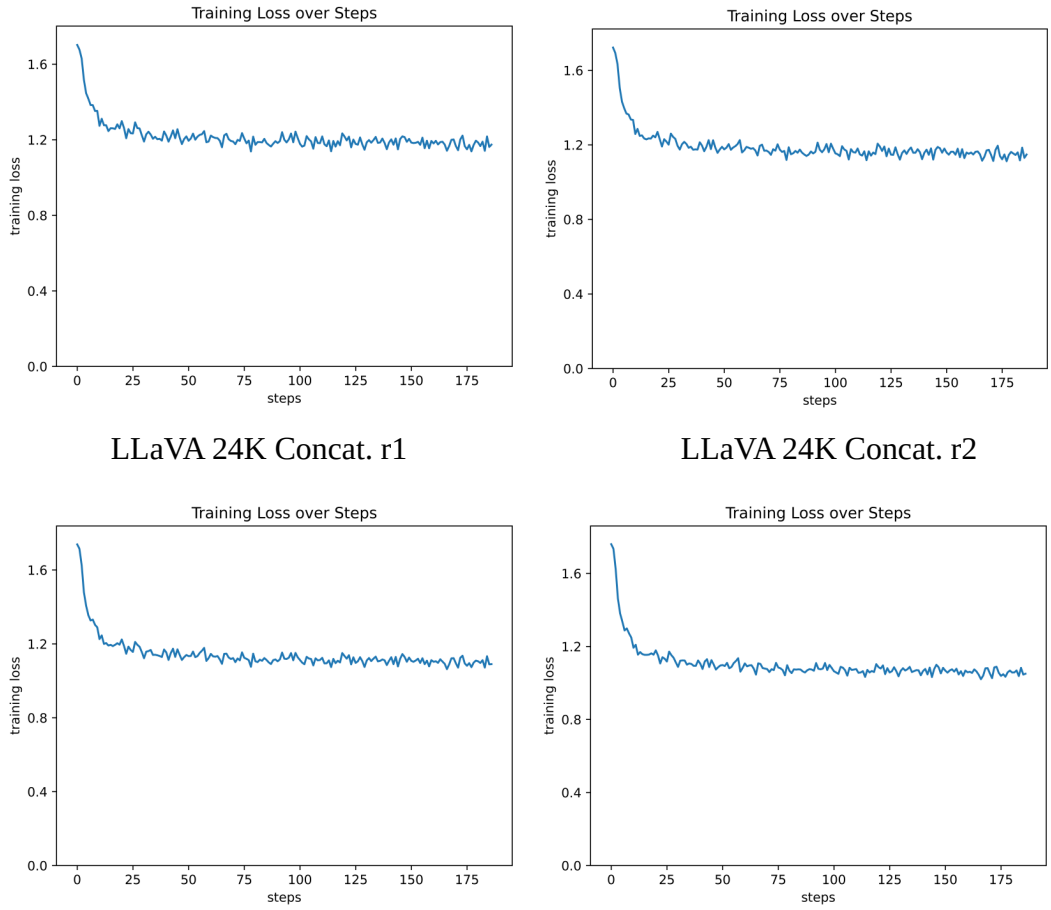


Figure 7. Visualization of fine-tuning loss of different conversation rounds.

Table 6. Ablation about only introducing negative samples on general benchmarks. Concat. and Neg. denotes concatenate and negative samples, respectively.

Dataset	MME^P	MME^C	MMB^T	MMB^D	POPE
LLaVA 24K	1039.62	265.36	41.48	42.53	82.51
LLaVA 24K Concat. Neg. (r1)	1304.77	268.21	40.87	41.87	81.65
LLaVA 24K Concat. Neg. (r2)	1336.41	258.57	24.74	18.21	80.77
LLaVA 24K Concat. Neg. (r4)	1304.25	259.29	25.00	28.52	82.42
LLaVA 24K Concat. Neg. (all)	1308.59	254.64	15.53	18.21	82.22

proportional representation. We explore various combination methods, such as replacing the original LLaVA data or directly adding MMVU data. Our results, presented in Tab. 13, reveal that adding new data does not consistently improve performance; replacement yields more substantial enhancements. The model achieves optimal performance when the replacement proportion closely matches the original distribution.

Table 7. Ablation about only introducing negative samples on the MMVU test set. Concat. and Neg. denotes concatenate and negative samples, respectively.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA↑
LLaVA 24K	17.50	22.73	20.83	12.50	45.83	16.67	13.64	50.00	28.57	29.03	45.83	31.82	27.33
LLaVA 24K Concat. Neg. (r1)	22.50	22.73	16.67	16.67	45.83	20.83	9.09	45.45	23.81	22.58	45.83	27.27	26.33
LLaVA 24K Concat. Neg. (r2)	17.50	13.64	8.33	12.50	58.33	16.67	9.09	45.45	38.10	16.13	37.50	31.82	24.67
LLaVA 24K Concat. Neg. (r4)	17.50	9.09	20.83	16.67	70.83	16.67	13.64	45.45	33.33	19.35	45.83	36.36	28.00
LLaVA 24K Concat. Neg. (all)	20.00	13.64	8.33	16.67	58.33	12.50	4.55	40.91	19.05	9.68	25.00	31.82	21.33
Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
LLaVA 24K	72.00	54.55	68.75	75.00	38.89	77.78	62.50	45.00	68.42	62.50	38.89	46.15	59.41
LLaVA 24K Concat. Neg. (r1)	65.38	61.54	73.33	69.23	42.11	73.68	80.00	47.37	73.68	72.00	38.89	53.85	62.20
LLaVA 24K Concat. Neg. (r2)	65.00	72.73	85.71	75.00	33.33	76.47	80.00	47.37	60.00	77.27	43.75	53.33	62.44
LLaVA 24K Concat. Neg. (r4)	63.16	80.00	66.67	60.00	19.05	75.00	66.67	47.37	65.00	71.43	42.11	42.86	56.48
LLaVA 24K Concat. Neg. (all)	52.94	62.50	85.71	63.64	26.32	80.00	87.50	50.00	77.78	83.33	53.85	50.00	63.01

Table 8. Ablation about different question types.

Dataset	MME^P	MME^C	MMB^T	MMB^D	POPE
Character	869.21	203.93	58.63	58.59	66.18
Semantic	1167.73	216.78	61.43	61.52	71.33
Understanding	918.38	219.64	60.15	59.88	67.93
Character + Semantic	963.81	225.36	59.98	60.05	68.84
Semantic + Understanding	1260.70	246.43	62.05	61.77	71.82
Character + Semantic + Understanding	1051.6	242.5	61.27	61.00	71.63

Table 9. Ablation about only introducing negative samples on the MMVU test set. Concat. and Neg. denotes concatenate and negative samples, respectively.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MA ↑
Character	35.00	40.91	20.83	20.83	62.50	33.33	22.73	59.09	33.33	25.81	50.00	31.82	36.00
Semantic	27.50	31.82	33.33	33.33	79.17	29.17	18.18	45.45	47.62	35.48	58.33	45.45	39.67
Understanding	47.50	40.91	29.17	33.33	66.67	41.67	31.82	54.55	47.62	38.71	58.33	40.91	44.33
Character + Semantic	45.00	40.91	29.17	37.50	66.67	33.33	27.27	59.09	47.62	35.48	54.17	40.91	43.00
Semantic + Understanding	32.50	45.45	33.33	41.67	70.83	20.83	18.18	68.18	28.57	38.71	58.33	45.45	41.33
Character + Semantic + Understanding	42.50	36.36	29.17	37.50	70.83	29.17	27.27	54.55	42.86	35.48	58.33	45.45	42.33
Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Character	46.15	35.71	61.54	66.67	21.05	60.00	44.44	38.10	58.82	68.00	29.41	50.00	48.57
Semantic	62.07	30.00	52.94	38.46	5.00	63.16	66.67	52.38	47.37	57.69	26.32	41.18	46.40
Understanding	26.92	30.77	53.33	33.33	20.00	54.55	30.00	42.86	44.44	53.85	26.32	43.75	38.99
Character + Semantic	30.77	35.71	58.82	25.00	23.81	57.89	45.45	38.10	47.37	57.69	27.78	47.06	41.63
Semantic + Understanding	51.85	23.08	52.94	28.57	15.00	72.22	63.64	31.82	68.42	55.56	30.00	44.44	45.13
Character + Semantic + Understanding	37.04	33.33	56.25	25.00	19.05	63.16	50.00	45.45	50.00	54.17	26.32	44.44	42.27

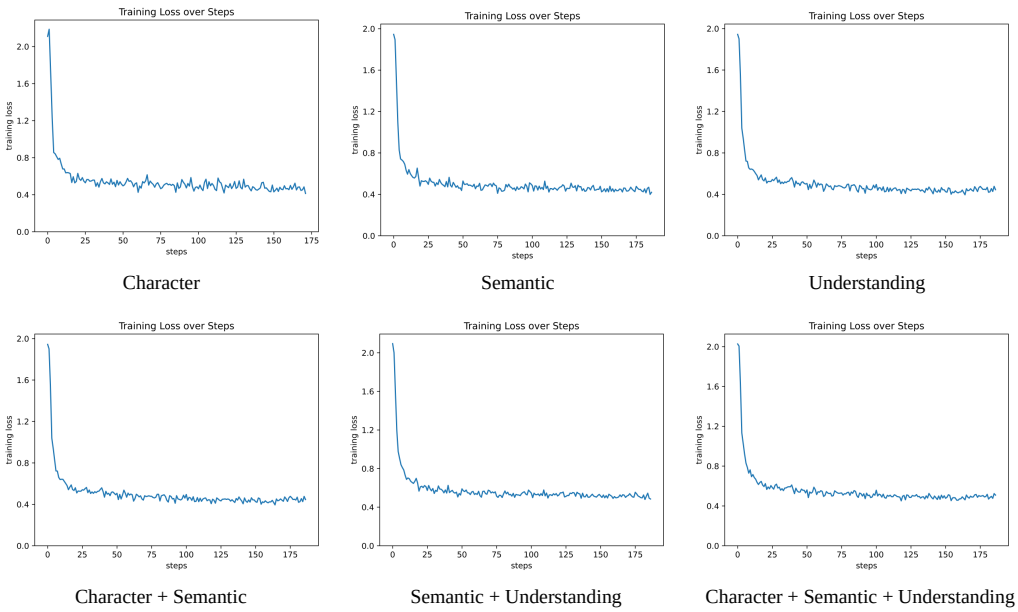


Figure 8. Visualization of fine-tuning loss of different data composition (version 1).

Table 10. Ablation about different combinations of data. "Version 2/3-Pos." denotes positive samples generated by the respective prompt version, while "Version 2/3-Neg." represents the corresponding negative samples.

Dataset	MME^P	MME^C	MMB^T	MMB^D	POPE
Version 2-Pos.	714.63	241.07	59.52	60.65	74.85
Version 2-Neg.	500.0	200.0	-	-	-
Version 2-Pos.+Neg.	1289.56	276.43	59.70	61.43	75.78
Version 3-Pos.	648.91	233.57	60.81	61.25	70.16
Version 3-Neg.	604.58	200.0	59.53	60.05	54.47
Version 3-Pos.+Neg.	1321.89	255.36	63.11	63.32	79.62

5. Examples of Paired Positive and Negative Samples Generated

STEP-2: Generate questions and corresponding correct answers based on the extracted information in STEP-1, Input Image, and Question Types (Do not output):

Question Types:

1. Ask if a certain object exists, the question is defined as ques-pos. Provide the correct answer to ques-pos as ans-pos. (2) Replace the object mentioned in the ques-pos and ans-pos with the visually similar nonexistent object. Ask if the model identifies the replaced object, the question is defined as ques-neg. Rhetorical questions are preferred. Provide the correct answer to ques-neg as ans-neg.
2. Ask a question about the name/category/color/texture/shape/pose of an object in the image, the question is defined as ques-pos. Provide the correct answer to ques-pos as ans-pos. (2) Replace the name/category/color/texture/shape/pose of the object mentioned in the ques-pos. Ask a question about if the modified name/category/color/texture/shape/pose of an object is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. Provide the correct answer to ques-neg as ans-neg.
3. Ask a question about the number/position of an object in the image, the question is defined as ques-pos. Provide the correct answer to ques-pos as ans-pos. (2) Replace the number/position of the object mentioned in the ques-pos with a similar number/position as interference. Ask a question about if the modified number/position of the object is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. Provide the correct answer to ques-neg as ans-neg.
4. Ask a question about the topic of abstract-knowledge/concrete-knowledge/professional-knowledge in the image, the question is defined as ques-pos. Provide the correct answer to ques-pos as ans-pos. (2) Replace the knowledge related to the topic mentioned in the ques-pos with similar knowledge as interference. Ask a distractor question about if the modified knowledge related to the topic is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. Provide the correct answer to ques-neg as ans-neg. - Candidate topics of abstract knowledge: emotions, aesthetics, connotations, symbols, culture, news, allusions, legends, common sense, functions, and so on. - Candidate topics of concrete knowledge: landmarks, celebrities, well-known objects, and so on. - Candidate topics of professional knowledge: specific knowledge in various vertical fields, industries, disciplines, and so on.
5. Ask a question about the activity of an object and/or interaction/relationship between objects or persons in the image, the question is defined as ques-pos. Provide the correct answer to ques-pos as ans-pos. (2) Replace the activity/interaction/relationship mentioned in the ques-pos with a similar activity/interaction/relationship as interference. Ask a distractor question about if the modified activity/interaction/relationship is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. Provide the correct answer to ques-neg as ans-neg.
6. NOTE: Pose questions that differ slightly from true understanding but are very similar, or add a distractor sentence to questions, so that the model is guided to make incorrect inferences or answers.
7. NOTE: Pose questions that differ slightly from true understanding but are very similar, or add a distractor sentence to questions, so that the model is guided to make incorrect inferences or answers.

Requirements:

1. Generate 7-10 questions.
2. Ask directly, without adding any information such as ”in the conversation,” ”mentioned in the conversation,” or ”in the image.”
3. Answer directly, with explanation, succinctly, with responses not exceeding 30 characters.
4. Questions must be real and answerable. Answers must be correct, definitely, and not speculative.
5. Do not generate questions that the answer is ”uncertain”.
6. The format of the generated questions and answers is: ’id’: ’json[’id’]’, ’image’: ’json[’image’]’, ’conversations-pos’: [’from’: ’human’, ’value’: ’ques-pos’, ’from’: ’gpt’, ’value’: ’ans-pos’,..., ’from’: ’human’, ’value’: ’ques-pos’, ’from’: ’gpt’, ’value’: ’ans-pos’], ’conversations-neg’: [’from’: ’human’, ’value’: ’ques-neg ’, ’from’: ’gpt’, ’value’: ’ans-neg’,..., ’from’: ’human’, ’value’: ’ques-neg’, ’from’: ’gpt’, ’value’: ’ans-neg’]

STEP-3: Output JSON sample

Figure 9. GPT4 prompt for *version 2*.

STEP-2: Generate questions and corresponding correct answers based on the extracted information in STEP-1, Input Image, and Question-Answer Types:

Question-Answer Types:

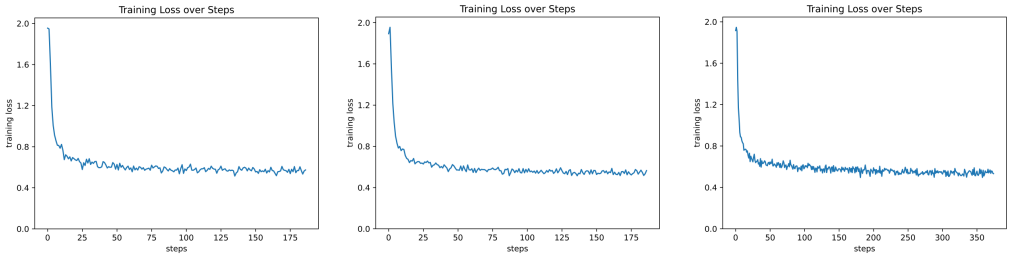
1. **Types-1:** (1) Ask if a certain object exists, the question is defined as ques-pos. (2) Operation-pos (3) Replace the object mentioned in the ques-pos with the visually similar nonexistent object. Ask if the model identifies the replaced object, the question is defined as ques-neg. Rhetorical questions are preferred. (4) Operation-neg
2. **Types-2:** (1) Ask a question about the name/category/color/texture/shape/pose of an object in the image, the question is defined as ques-pos. (2) Operation-pos (3) Replace the name/category/color/texture/shape/pose of the object mentioned in the ques-pos. Ask a question about whether the modified name/category/color/texture/shape/pose of an object is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. (4) Operation-neg
3. **Types-3:**(1) Ask a question about the number/position of an object in the image, the question is defined as ques-pos. (2) Operation-pos (3) Replace the number/position of the object mentioned in the ques-pos with a similar number/position as interference. Ask a question about if the modified number/position of the object is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. (4) Operation-neg
4. **Types-4:**(1) Ask a question about the topic of abstract-knowledge/concrete-knowledge/professional-knowledge in the image, the question is defined as ques-pos. (2) Operation-pos (3) Replace the knowledge related to the topic mentioned in the ques-pos with similar knowledge as interference. Ask a distractor question about if the modified knowledge related to the topic is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. (4) Operation-neg
   - Candidate topics of abstract knowledge: emotions, aesthetics, connotations, symbols, culture, news, allusions, legends, common sense, functions, and so on.
   - Candidate topics of concrete knowledge: landmarks, celebrities, well-known objects, and so on.
   - Candidate topics of professional knowledge: specific knowledge in various vertical fields, industries, disciplines, and so on.
5. **Types-5:**(1) Ask a question about the activity of an object and/or interaction/relationship between objects or persons in the image, the question is defined as ques-pos. (2) Operation-pos (3) Replace the activity/interaction/relationship mentioned in the ques-pos with a similar activity/interaction/relationship as interference. Ask a distractor question about if the modified activity/interaction/relationship is correct. Rhetorical questions are preferred. The new question is defined as ques-neg. (4) Operation-neg

Notes:

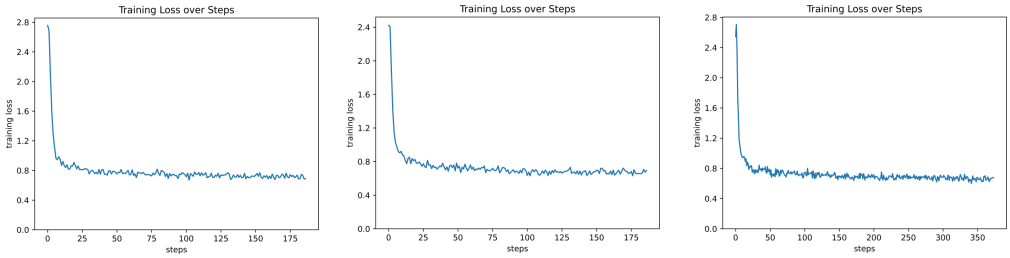
- All (2) Operation-pos is the same type: Design and provide four options based on ques-pos, including opt-pos-0: right answer with the correct reason, opt-pos-1: right answer with an incorrect reason, opt-pos-2: wrong answer with the correct reason, opt-pos-3: wrong answer with an incorrect reason. The opt-pos-0 is defined as ans-pos.
- All (4) Operation-neg is the same type: Design and provide four options based on ques-neg, including opt-neg-0: right answer with the correct reason, opt-neg-1: right answer with an incorrect reason, opt-neg-2: wrong answer with the correct reason, opt-neg-3: wrong answer with an incorrect reason. The opt-neg-0 is defined as ans-neg.
- The correct reason is the only one, while the incorrect reason can stem from various distortions and fabrications of facts, etc.
- Pose questions that differ slightly from true understanding but are very similar or add a distractor sentence to questions so that the model is guided to make incorrect inferences or answers.
- For ques-neg, rhetorical questions are preferred.

STEP-3: Output JSON sample

Figure 10. Instructions for generating questions and answers based on extracted image information. GPT4 prompt for *version 3*.



Version 2 – Pos. Version 2 – Neg. Version 2 – Pos. + Neg.



Version 3 – Pos. Version 3 – Neg. Version 3 – Pos. + Neg.

Figure 11. Visualization of fine-tuning loss of different data compositions (version 2 and version 3).

Table 11. Ablation about different combinations of data on MMVU test set. "Version 2/3-Pos." denotes positive samples generated by the respective prompt version, while "Version 2/3-Neg." represents the corresponding negative samples.

Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. RA↑
Version 2-Pos.	22.5	18.18	20.83	33.33	70.83	20.83	18.18	50.0	28.57	19.35	50.0	18.18	30.33
Version 2-Neg.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Version 2-Pos.+Neg.	20.0	9.09	20.83	33.33	58.33	25.0	13.64	45.45	23.81	19.35	50.0	18.18	27.67
Version 3-Pos.	20.0	18.18	16.67	16.67	58.33	20.83	18.18	45.45	23.81	19.35	50.0	27.27	27.33
Version 3-Neg.	27.5	9.09	16.67	8.33	45.83	20.83	0.0	45.45	23.81	32.26	45.83	31.82	26.0
Version 3-Pos.+Neg.	40.0	18.18	50.0	33.33	70.83	25.0	22.73	63.64	57.14	41.94	58.33	40.91	43.33
Method	Char/Num	Pres.	Color/Tex	Num.	Shape	Stance	Pos.	Abstract.	Concrete.	Expert.	Act.	Rel.	Avg. MR ↓
Version 2-Pos.	67.86	73.33	73.68	46.67	19.05	77.27	69.23	50.0	68.42	76.0	36.84	73.33	60.94
Version 2-Neg.	100.0	100.0	100.0	0	100.0	100.0	0	100.0	100.0	100.0	100.0	100.0	100.0
Version 2-Pos.+Neg.	72.41	80.0	70.59	50.0	36.36	71.43	66.67	52.38	73.68	76.0	42.86	75.0	63.27
Version 3-Pos.	70.37	69.23	75.0	71.43	33.33	77.27	66.67	52.38	75.0	75.0	40.0	62.5	63.72
Version 3-Neg.	50.0	75.0	69.23	80.0	38.89	66.67	100.0	52.38	68.75	52.38	35.29	36.36	56.18
Version 3-Pos.+Neg.	44.83	66.67	25.0	38.46	22.73	70.0	44.44	33.33	33.33	45.83	30.0	43.75	40.91

Table 12. Comparison to leading MLLMs on 10 benchmarks. [§]We evaluate the officially released checkpoint by ourselves.

Model	Vision Encoder	LLM	MME^{P}	MME^{C}	MMB^{T}	MMB^{D}	SEED	$MMMU^{V}$	$MMMU^{T}$	VQA^{v2}	GQA	POPE
IDEFICS-80B [16]	OpenCLIP-H (1.0B)	LLaMA-65B	–	–	54.6	54.5	–	–	–	60.0	–	–
BLIP-2 [21]	EVA01-CLIP-G (1.0B)	Vicuna-13B	–	–	–	–	–	–	–	–	41.0	–
InstructBLIP [9]	EVA01-CLIP-G (1.0B)	Vicuna-13B	–	–	–	–	–	–	–	–	49.5	83.7
BLIP-2 [21]	EVA01-CLIP-G (1.0B)	Flan-T5-XXL (11B)	1293.8	290.0	–	–	–	35.4	34.0	65.0	44.6	–
InstructBLIP [9]	EVA01-CLIP-G (1.0B)	Flan-T5-XXL (11B)	1212.8	291.8	–	–	–	35.7	33.8	–	47.9	–
Shikra-13B [4]	CLIP-L (0.4B)	Vicuna-13B	–	–	–	–	–	–	–	77.4	–	–
LLaVA-v1.5-13B (LoRA) [25]	CLIP-L (0.4B)	Vicuna-13B	1541.7	300.4[§]	68.4[§]	68.5	61.3	40.0[§]	33.2[§]	80.0	63.3	86.7
VILA-13B [24]	CLIP-L (0.4B)	LLaMA2-13B	1570.1	-	70.3	-	62.8	-	-	80.8	63.3	84.2
SPHINX-Plus [11]	Mixture of Visual Experts (MoV)	LLaMA2-13B	1457.7	283.6	71.0	-	74.8	-	-	-	-	89.1
LLaVA-Llama-3-8B-v1.1 [8]	CLIP-L (0.4B)	Llama-3-8B	1469	349	72.3	-	-	36.8	-	-	62.6	86.4
InstructBLIP [9]	EVA01-CLIP-G (1.0B)	Vicuna-7B	–	–	33.9	36.0	53.4	–	–	–	49.2	–
MiniGPT-v2 [3]	EVA01-CLIP-G (1.0B)	LLaMA2-7B	–	–	–	–	–	–	–	–	60.3	–
IDEFICS-9B [16]	OpenCLIP-H (1.0B)	LLaMA-7B	–	–	45.3	48.2	–	–	–	50.9	–	–
LLaVA-v1.5-7B (LoRA) [25]	CLIP-L (0.4B)	Vicuna-7B	1476.9	267.9[§]	66.1[§]	66.1	60.1	34.4[§]	31.7[§]	79.1	63.0	86.4
mPLUG-Owl2 [34]	CLIP-L (0.4B)	LLaMA2-7B	1450.2	313.2	66.0	66.5	57.8	32.7	32.1	79.4	56.1	85.8
VILA-7B [24]	CLIP-L (0.4B)	LLaMA2-7B	1533.0	-	68.9	-	61.1	-	-	79.9	62.3	85.5
Shikra-7B [4]	CLIP-L (0.4B)	Vicuna-7B	–	–	60.2	58.8	–	–	–	–	–	–
SPHINX-Intern2 [11]	Mixture of Visual Experts (MoV)	InternLM2-7B	1260.4	294.6	57.9	-	68.8	-	-	75.5	56.2	86.9
Mini-Gemini [23]	CLIP	Vicuna-7B	1523	316	-	-	-	36.1	32.8	65.2	-	-
MM1-7B-Chat [28]	ViT-H (5B)	MM1-7B	1529.3	328.9	72.3	-	64.0	37.0	35.6	82.8	-	86.6
LLaVA-Phi-3-mini [8]	CLIP-L (0.4B)	Phi-3 (4B)	1477	313	69.2	-	-	41.4	-	-	61.5	87.3
Gemini Nano-2 [32]	–	Gemini Nano-2 (3.25B)	–	–	–	–	–	32.6	–	67.5	–	–
MM1-3B-Chat [28]	ViT-H (5B)	MM1-3B	1482.5	279.3	67.8	-	63.0	33.9	33.7	82.0	-	87.4
MobileVLM [6]	CLIP-L (0.4B)	MobileLLaMA (2.7B)	1288.9	–	–	59.6	–	–	–	–	59.0	84.9
MobileVLM V2 [7]	CLIP-L (0.4B)	MobileLLaMA (2.7B)	1440.5	–	–	63.2	–	–	–	–	61.1	84.7
ALLaVA-Longer [2]	SigLIP-SO (0.4B)	Phi-2 (2.7B)	-	-	64.6	-	65.6	33.2	-	-	50.0	-
TinyLLaVA-share-Sig-Phi [38]	SigLIP-SO (0.4B)	Phi-2 (2.7B)	1464.9	-	66.9	-	-	-	-	79.9	62.0	86.4
Bunny-MMVU-3B	SigLIP-SO (0.4B)	Phi-2 (2.7B)	1515.1	262.9	70.7	69.7	63.2	38.2	34.0	80.1	62.1	86.0
Bunny-MMVU-4B	SigLIP-SO (0.4B)	Phi-3 (4B)	1511.2	313.2	72.3	71.8	64.1	40.9	38.8	80.4	62.2	86.3
Bunny-MMVU-8B	SigLIP-SO (0.4B)	Llama-3-8B	1606.1	331.8	77.6	76.3	65.1	42.6	38.7	81.5	63.4	86.3

Table 13. Ablation about the combination of the proposed MMVU dataset with other datasets. We train these datasets with the same vision encoder (SigLIP).

Dataset	LLM	$MMVU^{A}\uparrow$	$MMVU^{M}\downarrow$	MME^{P}	MME^{C}	MMB^{T}	MMB^{D}	SEED	$MMMU^{V}$	$MMMU^{T}$	VQA^{v2}	GQA	POPE
Mix 106k	Phi-2	44.00	43.59	1410.01	266.07	66.14	66.15	60.07	36.70	32.90	75.71	56.39	84.89
Mix 106k-replace MMVU-Train 12k	Phi-2	45.33	41.63	1432.51	271.78	61.43	60.91	57.97	35.00	33.20	75.75	56.50	83.64
Mix 106k-replace MMVU-Train 24k	Phi-2	47.00	40.51	1421.54	248.57	67.26	66.32	60.37	38.00	34.00	75.95	56.92	84.40
Mix 106k-replace MMVU-Train 48k	Phi-2	47.33	38.26	1419.26	239.29	66.76	65.64	60.61	35.60	33.70	76.24	56.96	83.63
Mix 106k + MMVU-Train 12k	Phi-2	50.00	36.71	1388.89	252.5	65.86	66.07	60.51	37.30	33.00	75.86	56.87	84.06
Mix 106k + MMVU-Train 24k	Phi-2	48.67	38.66	1400.74	244.64	66.98	65.98	60.62	37.10	32.90	76.05	56.96	83.73
Mix 106k + MMVU-Train 48k	Phi-2	47.00	38.16	1410.40	247.5	65.86	66.24	60.85	36.80	33.90	73.45	57.05	84.98
Base 695k	Phi-2	55.33	27.19	1492.17	274.29	68.27	68.56	62.31	37.40	32.80	79.85	61.84	86.52
Base-replace MMVU-Train 112k (649k)	Phi-2	57.00	22.97	1498.57	255.35	70.35	68.90	62.88	35.70	33.40	79.69	62.08	85.13
Weight merge	Phi-2	58.33	23.91	1515.06	262.86	70.74	69.67	63.23	38.20	34.00	80.14	62.10	85.98
Base 695k	Phi-3	54.67	29.61	1462.07	335.35	70.57	71.56	63.08	40.30	39.00	80.34	61.69	85.21
Base-replace MMVU-Train 112k (649k)	Phi-3	58.33	24.57	1520.43	283.93	72.43	71.23	63.89	41.30	38.50	80.14	61.74	85.88
Weight merge	Phi-3	58.33	27.39	1511.16	313.2	72.31	71.82	64.12	40.90	38.80	80.44	62.21	86.28
Base 695k	Llama-3-8B	58.00	26.27	1548.53	313.57	76.63	75.34	64.62	40.70	38.60	80.99	62.82	85.30
Base-replace MMVU-Train 112k (649k)	Llama-3-8B	61.00	21.12	1600.46	318.93	75.22	75.26	64.62	41.70	37.90	81.03	62.72	86.30
Weight merge	Llama-3-8B	59.33	24.58	1606.12	331.79	77.63	76.29	65.11	42.60	38.70	81.52	63.44	86.25

Conversation

Question: Are the people preparing for a party?
Answer: No, they are preparing a meal.

Question: Is there a green chair visible?
Answer: No, no green chair.

Question: Are there three people in the kitchen?
Answer: No, only two people.

Question: Is the kitchen light pink?
Answer: No, it is not pink.

Question: Does the window show a garden outside?
Answer: No visible garden.

Figure 12. Generated conversation example by prompt version 0.

Conversation (Question type: Character)

Question: Is there a number 358 visible anywhere?
Answer: No, there are no numbers visible.

Question: Does the text 'Happy Birthday' appear in the kitchen?
Answer: No, there is no text visible.

Conversation (Question type: Semantic)

Question: Is there a microwave in the image?
Answer: No, there is no microwave visible.

Question: Can you see a refrigerator in the scene?
Answer: No, there is no refrigerator visible.

Question: Is there a blender on the counter?
Answer: No, there is no blender shown.

Question: Are there five plates in the picture?
Answer: No, there are more than five plates visible.

Conversation (Question type: Understanding)

Question: Are the individuals preparing for a meal?
Answer: Yes, serving food.

Question: Is the setting a restaurant kitchen?
Answer: No, it's a home kitchen.

Figure 13. Generated conversation example by prompt version 1.

Positive Conversation

Question: Is there a bowl of food on the counter?
Answer: Yes, there are multiple bowls of food on the counter.

Question: What is the man wearing?
Answer: The man is wearing a plaid shirt.

Question: How many people are in the kitchen?
Answer: There are two people in the kitchen.

Question: What are the people doing?
Answer: They are preparing a meal.

Question: Where is the sink located in the kitchen?
Answer: The sink is located towards the right.

Negative Conversation

Question: Is there a vase of flowers on the counter?
Answer: No, there is no vase of flowers on the counter.

Question: What is the man wearing?
Answer: The man is wearing a green hoodie.

Question: How many children are in the kitchen?
Answer: There are no children in the kitchen.

Question: What are some people doing?
Answer: They are reading books.

Question: Where is the dishwasher located in the kitchen?
Answer: The dishwasher is located towards the right.

Figure 14. Generated conversation example by prompt version 2.

Positive Conversation

Question: Are there two people in the kitchen?
Option: {"Yes, there are two people.", "Yes, there are three people.", "No, there is only one person.", "No, there are four people."}
Answer: Yes, there are two people.

Question: Are the people in the kitchen cooking?
Option: {"Yes, they are cooking.", "No, they are cleaning.", "No, they are eating.", "Yes, they are fixing something."}
Answer: Yes, they are cooking.

Question: How many bowls are there on the counter?
Option: {"One bowl.", "Four bowls.", "Six bowls.", "Two bowls."}
Answer: Four bowls.

Question: What is the main activity happening?
Option: {"Preparing a meal.", "Washing dishes.", "Playing games.", "Reading books."}
Answer: Preparing a meal.

Question: Is there a dishwasher in the image?
Option: {"Yes, the dishwasher is there.", "No, there's no dishwasher.", "It's unclear.", "There is a microwave instead."}
Answer: Yes, the dishwasher is there.

Question: Is there a pot on the table?
Option: {"Yes, there is a pot.", "No, there isn't a pot.", "Yes, there are two pots.", "No, there are only pans."}
Answer: Yes, there is a pot.

Question: Are the people preparing food in a modern kitchen?
Option: {"Yes, they are.", "No, it's an old kitchen.", "Yes, in a living room.", "No, it's a commercial kitchen."}
Answer: Yes, they are.

Figure 15. Generated conversation example by prompt version 3 (part 1).

Negative Conversation

Question: Are there three people in the kitchen?
Option: {"Yes, there are two people.", "Yes, there are three people.", "No, only one person.", "No, four people."}
Answer: No, only one person.

Question: Are the people in the kitchen cleaning?
Option: {"Yes, they are cooking.", "No, they are cleaning.", "No, they are eating.", "Yes, fixing something."}
Answer: Yes, they are cooking.

Question: How many bowls are there on the countertop?
Option: {"One bowl.", "Four bowls.", "Six bowls.", "Two bowls."}
Answer: Four bowls.

Question: What is the secondary activity happening?
Option: {"Preparing a meal.", "Washing dishes.", "Playing games.", "Reading books."}
Answer: Preparing a meal.

Question: Is there a microwave in the image?
Option: {"Yes, the microwave is there.", "No, there's no microwave.", "It's unclear.", "There is a dishwasher."}
Answer: There is a dishwasher.

Question: Is there a pan on the table?
Option: {"Yes, there is a pot.", "No, there isn't a pot.", "Yes, there are two pans.", "No, only bowls."}
Answer: Yes, there is a pot.

Question: Are the people preparing food in a commercial kitchen?
Option: {"Yes, they are.", "No, it's old kitchen.", "Yes, a living room.", "No, modern kitchen."}
Answer: No, modern kitchen.

Figure 16. Generated conversation example by prompt version 3 (part 2).

6. The Details of "Instruction prompt".

7. Details about Content Guided Refinement Strategy

7.1. Prompt for extraction visual information

The prompt is shown in Fig. 18.

7.2. Examples of Content Guided Refinement Strategy Results

The example is shown in Fig. 19.

8. Details about Visual Attention Refinement Strategy

8.1. Detailed Description

Our analysis indicates that the probability of the model generating correct answers decreases when the attention values between the question token and the visual token are relatively low. To address this, we propose leveraging the attention mechanism between the question and visual tokens to optimize the visual prompts, as shown in the algorithm 1. Specifically, we extract the attention parameters from the final transformer layer of the MLLM, while other proposed steps remain consistent with those outlined in the main text. Using these parameters, an attention map is generated, where regions with higher values indicate areas of focus for the model. To guide the model toward regions relevant to the question, we invert the attention map, apply regularization and filtering, and blend it with the original image. The blending process employs coefficients of 0.85 for the original image and 0.15 for the attention map.

8.2. Examples of visual attention refinement strategy

The example is shown in Fig. 20.

9. Broader Impact and Limitation

Broader impact. In this paper, we study how to evaluate and improve MLLMs' robustness when answering challenging visual questions. The research will benefit the general development and applications of MLLMs.

Limitation. Due to the cost of dataset construction, we contribute 112K instruction tuning data. Such adversarial datasets can be scaled up with more quota.

Algorithm 1 Optimizing Visual Prompts Using Attention Mechanisms

Require: MLLM with pre-trained parameters, input question Q, input image I, blending coefficients $\alpha = 0.85$, $\beta = 0.15$

Ensure: Optimized visual prompt I' for inference

1: Extract attention parameters A from the final transformer layer of the MLLM.
2: Compute the attention map M using A between the question token and the visual tokens.
3: Identify high-value regions in M where the model focuses during inference.
4: **Invert the attention map:**
5: Compute the inverted attention map $M_{\text{inv}} = 1 - M$.
6: **Regularize and filter:**
7: Apply normalization and filtering to M_{inv} to obtain M_{filtered}.
8: **Blend attention map with the original image:**
9: Generate the optimized image I' as:

$$I' = \alpha \cdot I + \beta \cdot M_{\text{filtered}}$$

10: **return** Optimized image I' for inference.

STEP-1 (DO NOT OUTPUT): Image Information Extraction

1. **Text/Numbers:** Identify clear text or numbers in the image and their positions.
2. **Objects:** Identify objects, their properties (name/category/color/texture/shape/pose), position, and quantity.
3. **People:** Identify people, and their attributes (name, position, gender, facial expression), and distinguish between real people and portraits in photos or posters.
4. **Object Relationships:** Understand relationships between objects (relative positions or hierarchy).
5. **People-Object Relationships:** Understand relationships between people and objects (e.g., someone riding a bike).
6. **Activities:** If there are multiple people, identify their activities.
7. **Events:** Understand events in the image and their cultural, emotional, or knowledge-related context.

STEP-2 (DO NOT OUTPUT): Carefully review the image and the question to ensure an accurate understanding of each question. Example Responses:

- **Question:** Is there a red bicycle in the image?
  Answer: Yes, there is a red bicycle in the image. It is clearly shown in the center.
- **Question:** Is the car in the image green?
  Answer: No, the car in the image is blue. The color is distinct.

STEP-3: Answer with the option's letter from the given choices directly.

Figure 17. Common Prompt for information extraction.

Extract information from the image and structure the output as follows:

1. **Text/Numbers:** Identify any visible text or numbers, including their content, font, and precise positions in the image.
2. **Objects:** Identify each object, detailing its name/category, color, texture, shape, pose, relative location within the image, and quantity.
3. **People:** Identify all visible individuals, with attributes such as position, gender, facial expression, and distinguish between real individuals and representations (e.g., in portraits, posters).
4. **Relationships & Interactions:** Describe spatial or contextual relationships between objects, and interactions or spatial relationships between people and objects.
5. **Activities:** Identify and describe any observed activities or actions involving people or objects.
6. **Events:** Identify any significant events or scenes occurring in the image.

Figure 18. Prompt for information extraction.

Example

ID: attribute/existence/COCO_val2014_000000170365
Image: COCO_val2014_000000170365.jpg
Question: What is the object under a plate of yellow plants in the picture?
Options:
A. Towel
B. Shower enclosure
C. Sink
D. Closestool
Ground Truth: D
Question Type: POS
Info:

1. **Text/Numbers:** There is no visible text or numbers in the image.
2. **Objects:**
   (a) Sink: A white porcelain sink with a silver faucet.
   (b) Toilet: A white porcelain toilet.
   (c) Bathtub: A white porcelain bathtub with a shower door.
   (d) Mirror: A framed mirror on the wall above the sink.
   (e) Potted Plant: A yellow plant in a pot on the counter.
   (f) Towels: White towels hanging on the towel rack.
   (g) Toilet Paper: A roll of toilet paper on the back of the toilet.
   (h) Tissue Box: A box of tissues on the counter.
   (i) Trash Can: A white trash can with a lid.
   (j) Toilet Brush: A white toilet brush in a holder.
3. **Relationships & Interactions:** Various spatial and functional relationships between objects, as described above.
4. **Activities:** None observed.
5. **Events:** The image shows a typical bathroom setup.

Figure 19. Examples of Content Guided Refinement Strategy Results.

POS Example

ID: attribute/existence/21
Image: 21.jpg
Question: How many puppies are there in the picture?
Options:
A. There is only one puppy in the picture and the other one is a cat.
B. There are two puppies in the picture.
C. There are no puppies in the picture, just two kittens.
D. There is no puppy in the picture, just two stuffed animals.
Ground Truth: A
Question Type: POS
Question: Does the picture show a dog and its pups?
Options:
A. The pup lay on top of the dog.
B. The picture shows a dog and its two puppies.
C. There are no puppies in the picture, just two kittens.
D. There is a dog in the picture and the other smaller animal is a cat.
Ground Truth: D
Question Type: NEG
Info:

1. **Text/Numbers:** There is no visible text or numbers in the image.
2. **Objects:**
   (a) Dog: White and brown, smooth fur, oval body, sitting on a blanket.
   (b) Cat: White and gray, soft fur, pointed ears, sitting on a blanket.
   (c) Blanket: White, soft, fluffy, laying on a wooden surface.
   (d) Wooden Surface: White, rough and uneven, under the blanket.
3. **People:** No visible people.
4. **Identification:** No real individuals, likely a portrait or poster.
5. **Activities:** No observed actions.
6. **Events:** No significant events.

Figure 20. Examples of visual attention refinement strategy.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, and Ahmed Awadallah et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 3, 4

[2] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 16

[3] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 16

[4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 16

[5] 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. 2

[6] Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*, 2023. 16

[7] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024. 16

[8] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner. 16

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 16

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 9

[11] Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024. 16

[12] 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. 8, 9

[13] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 3, 4, 7, 8

[14] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 8

[15] 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. 8, 9

[16] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. https://huggingface.co/blog/idefics, 2023. 16

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 8

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 8

[19] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 3, 4

[20] 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. 9

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 16

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, 2023. Association for Computational Linguistics. 9

[23] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 16

[24] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023. 16

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 16

[26] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 9

[27] 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. 8

[28] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 16

[29] Microsoft. Phi-2: The surprising power of small language models, 2023. 7

[30] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 8

[31] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 8

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 16

[33] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2023. 8

[34] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 16

[35] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 4

[36] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 9

[37] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 7

[38] Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024. 16