RLAIF-V: Open-Source AI Feedback Leads to Super GPT-4V Trustworthiness

Introduction

- 문제
    - 많은 Vision Language Model은 Properietary MLLM(ex: GPT-4V)에서 생성된 토큰을 Distillation하는 방식을 취하고 있음
        - 이는 단순히 Properietary MLLM과의 성능을 따라갈수만 있을 뿐, 넘어설 수는 없음
    - Scaling Laws in Test-time(CoT, Reasoning 등)
        - LLM에서 볼 수 있듯이 Test-time에서 Scaling Laws를 적용하면, 성능이 올라가는 것을 알 수 있음
        - 대부분의 Vision Language Model에서는 이걸 안하고 있음
            - 최근에 LLaVA-o1(2024.11)이 있긴하지만, 이 논문은 2024.05에 나왔음

Introduction

- 제안
    - novel deconfounded candidate response generation strategy를 제안
        - divide-and-conquer 접근을 사용
        - 같은 프롬프트를 주고 여러 개의 후보 응답을 생성하게 하고
          response 내에서 진정한 신뢰성 차이를 얻을 수 있음(GRPO 같은 느낌)
    - Inference time scaling guidance, novel self-feedback 접근
        - DPO를 사용하면 선호도를 맞추기 위해서 response의 길이가 짧아지는 것을 볼 수 있음
        - length-normalize strategy를 사용해서 token을 압축하려는(response의 길이가 짧아지는 것)
          을 방지할 수 있음

RLAIF-V

- Response Generation
    - 일반적인 방법
        1. Preference Learning을 위한 Feedback는 comparison pair(비교 쌍)의 형식으로 수집
        2. y_w(선호), y_l(비선호)로 수집되며 선호도를 학습
    - 일반적인 방법으로 학습하면 오로지 선호,비선호로만 학습이 되며 두 비교 쌍에 정답이 없을 수도 있고,
      샘플링된 두 개의 비교 쌍의 품질이 모두 안 좋을 수 있음
    - 제안하는 방법
        1. Preference Learning을 위한 비교 쌍이 두 개로 이루어지는 것이 아닌,
           많이 샘플링해서 수집(y_1, y_2, … y_n)

RLAIF-V

- Feedback Annotation
    - Divide
        1. 고정된 프롬프트에서 얻은 Response는 각각 여러 개의 문장으로 존재하고 있음
        2. 여러 개의 문장으로 존재하는 Response를 LLM을 이용해서 여러 개의 평서문 문장으로 분할
    - Conquer
        1. 분할된 문장이 맞는지 평가하기 위해서 이를 예, 아니오로 답변할 수 있는 질문으로 바꿈
        2. 이 질문에 대해서 다른 오픈소스 MLLM의 예 토큰과 아니오 토큰 둘 중 하나로 응답하도록 하고
           예 토큰의 probability를 선호도로 설정
    - Combine
        1. 분할된 질문에 대한 전체 선호도를 평균내서 전체 응답의 선호도로 설정
        2. 만약에 아니오 토큰으로 응답하는 질문이 더 많으면 -n_rej를 선호도로 설정
        3. 학습 비용을 줄이기 위해서 만들어진 하나의 instruction에서 최대 2개를 무작위로 샘플링해서 학습
        4. 너무 짧은 응답은 제외하고, y_w, y_l|이 하나의 토큰 이하가 되도록 샘플링

RLAIF-V

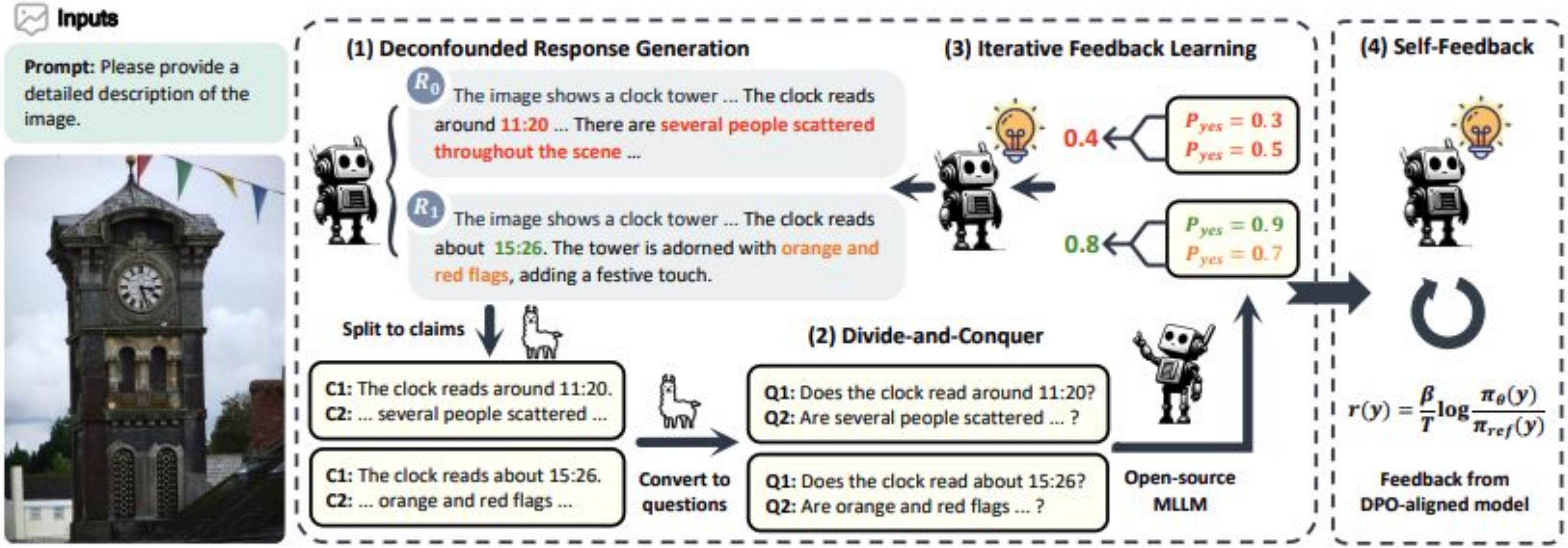
- Iterative Feedback Learning
  - 처음 만들어놓고 계속 사용하는 것이 아니라 특정 checkpoint마다 response 생성을 최신 checkpoint로 바꾸고 Divide, Conquer, Combine을 반복
    1. 각 반복을 시작할 때 N개의 instruction을 선택
    2. 최신 checkpoint로 Divide, Conquer, Combine을 수행
    3. 만들어진 Divide, Conquer, Combine으로 최신 checkpoint를 학습
    4. 1 ~ 3을 반복

RLAIF-V

- Self-Feedback for Inference-time Scaling
  - 그냥 DPO를 수행하면 점점 응답이 짧아지는 것을 확인
  - 이를 방지하기 위해 모든 토큰 수준의 score를 평균화해서 bias를 해결

$$r(y) = \beta \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)} = \beta \sum_{t}^{T} \log \frac{\pi_\theta(y_t|y_{<t})}{\pi_{\text{ref}}(y_t|y_{<t})}, \longrightarrow r(y) = \frac{\beta}{T} \log \frac{\pi_\theta(y)}{\pi_{\text{ref}}(y)}$$

Figure



Inputs

Prompt: Please provide a detailed description of the image.

(1) Deconfounded Response Generation

R_0 The image shows a clock tower ... The clock reads around 11:20 ... There are several people scattered throughout the scene ...

R_1 The image shows a clock tower ... The clock reads about 15:26. The tower is adorned with orange and red flags, adding a festive touch.

Split to claims

C1: The clock reads around 11:20.
C2: ... several people scattered ...

C1: The clock reads about 15:26.
C2: ... orange and red flags ...

Convert to questions

(2) Divide-and-Conquer

Q1: Does the clock read around 11:20?
Q2: Are several people scattered ... ?

Q1: Does the clock read about 15:26?
Q2: Are orange and red flags ... ?

Open-source MLLM

(3) Iterative Feedback Learning

0.4 ← $P_{yes} = 0.3$, $P_{yes} = 0.5$

0.8 ← $P_{yes} = 0.9$, $P_{yes} = 0.7$

(4) Self-Feedback

$$r(y) = \frac{\beta}{T}\log\frac{\pi_\theta(y)}{\pi_{ref}(y)}$$

Feedback from DPO-aligned model

Experiment

- Model
  - 첫 번째 모델
    1. use LLaVA 1.5 as the instruction model
    2. LLaVA-NeXT as labeler model
  - 두 번째 모델
    1. OmniLMM as instruction and labeler model

Experiment

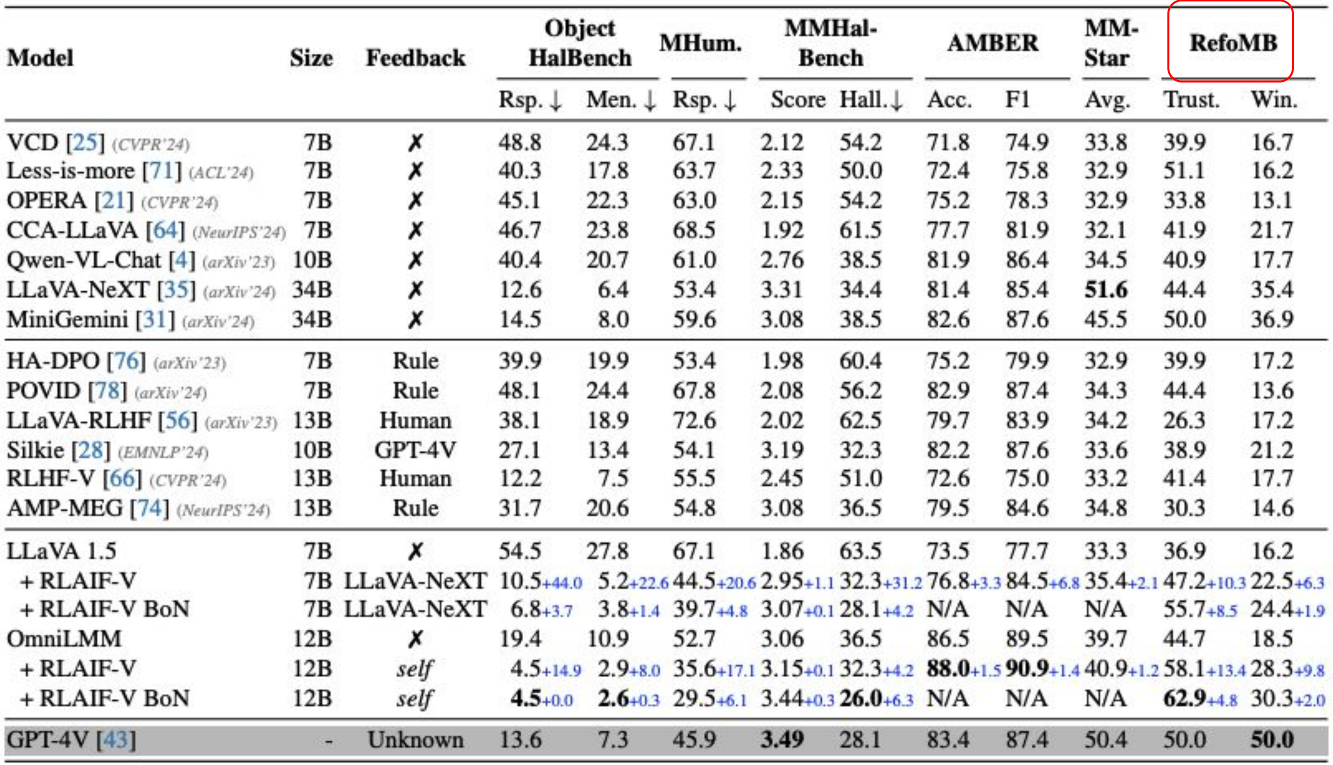
- Training Data
  - MSCOCO
  - ShareGPT-4V
  - MovieNet
  - Google Landmark v2
  - VQA v2
  - OKVQA
  - TextVQA

Experiment

- Evaluation
  - trustworthiness
    1. Object HalBench
    2. MMHal-Bench
    3. MHumanEval
    4. AMBERT
    5. MMStar

Results

이 논문에서 만든 벤치마크



Model	**Size**	**Feedback**	**Object HalBench**		**MHum.**	**MMHal-Bench**		**AMBER**		**MM-Star**	**RefoMB**	
			Rsp. ↓	Men. ↓	Rsp. ↓	Score	Hall.↓	Acc.	F1	Avg.	Trust.	Win.
VCD [25] *(CVPR'24)*	7B	✗	48.8	24.3	67.1	2.12	54.2	71.8	74.9	33.8	39.9	16.7
Less-is-more [71] *(ACL'24)*	7B	✗	40.3	17.8	63.7	2.33	50.0	72.4	75.8	32.9	51.1	16.2
OPERA [21] *(CVPR'24)*	7B	✗	45.1	22.3	63.0	2.15	54.2	75.2	78.3	32.9	33.8	13.1
CCA-LLaVA [64] *(NeurIPS'24)*	7B	✗	46.7	23.8	68.5	1.92	61.5	77.7	81.9	32.1	41.9	21.7
Qwen-VL-Chat [4] *(arXiv'23)*	10B	✗	40.4	20.7	61.0	2.76	38.5	81.9	86.4	34.5	40.9	17.7
LLaVA-NeXT [35] *(arXiv'24)*	34B	✗	12.6	6.4	53.4	3.31	34.4	81.4	85.4	**51.6**	44.4	35.4
MiniGemini [31] *(arXiv'24)*	34B	✗	14.5	8.0	59.6	3.08	38.5	82.6	87.6	45.5	50.0	36.9
HA-DPO [76] *(arXiv'23)*	7B	Rule	39.9	19.9	53.4	1.98	60.4	75.2	79.9	32.9	39.9	17.2
POVID [78] *(arXiv'24)*	7B	Rule	48.1	24.4	67.8	2.08	56.2	82.9	87.4	34.3	44.4	13.6
LLaVA-RLHF [56] *(arXiv'23)*	13B	Human	38.1	18.9	72.6	2.02	62.5	79.7	83.9	34.2	26.3	17.2
Silkie [28] *(EMNLP'24)*	10B	GPT-4V	27.1	13.4	54.1	3.19	32.3	82.2	87.6	33.6	38.9	21.2
RLHF-V [66] *(CVPR'24)*	13B	Human	12.2	7.5	55.5	2.45	51.0	72.6	75.0	33.2	41.4	17.7
AMP-MEG [74] *(NeurIPS'24)*	13B	Rule	31.7	20.6	54.8	3.08	36.5	79.5	84.6	34.8	30.3	14.6
LLaVA 1.5	7B	✗	54.5	27.8	67.1	1.86	63.5	73.5	77.7	33.3	36.9	16.2
+ RLAIF-V	7B	LLaVA-NeXT	10.5$_{+44.0}$	5.2$_{+22.6}$	44.5$_{+20.6}$	2.95$_{+1.1}$	32.3$_{+31.2}$	76.8$_{+3.3}$	84.5$_{+6.8}$	35.4$_{+2.1}$	47.2$_{+10.3}$	22.5$_{+6.3}$
+ RLAIF-V BoN	7B	LLaVA-NeXT	6.8$_{+3.7}$	3.8$_{+1.4}$	39.7$_{+4.8}$	3.07$_{+0.1}$	28.1$_{+4.2}$	N/A	N/A	N/A	55.7$_{+8.5}$	24.4$_{+1.9}$
OmniLMM	12B	✗	19.4	10.9	52.7	3.06	36.5	86.5	89.5	39.7	44.7	18.5
+ RLAIF-V	12B	*self*	4.5$_{+14.9}$	2.9$_{+8.0}$	35.6$_{+17.1}$	3.15$_{+0.1}$	32.3$_{+4.2}$	**88.0**$_{+1.5}$	**90.9**$_{+1.4}$	40.9$_{+1.2}$	58.1$_{+13.4}$	28.3$_{+9.8}$
+ RLAIF-V BoN	12B	*self*	**4.5**$_{+0.0}$	**2.6**$_{+0.3}$	29.5$_{+6.1}$	3.44$_{+0.3}$	**26.0**$_{+6.3}$	N/A	N/A	N/A	**62.9**$_{+4.8}$	30.3$_{+2.0}$
GPT-4V [43]	-	Unknown	13.6	7.3	45.9	**3.49**	28.1	83.4	87.4	50.4	50.0	**50.0**

Results

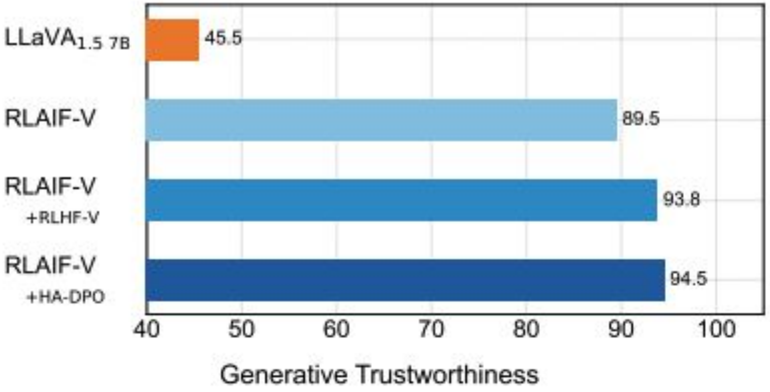


Figure 3. Results of combining RLAIF-V with other feedback. We report the response-level no-hallucination rate on Object HalBench for generative trustworthiness.

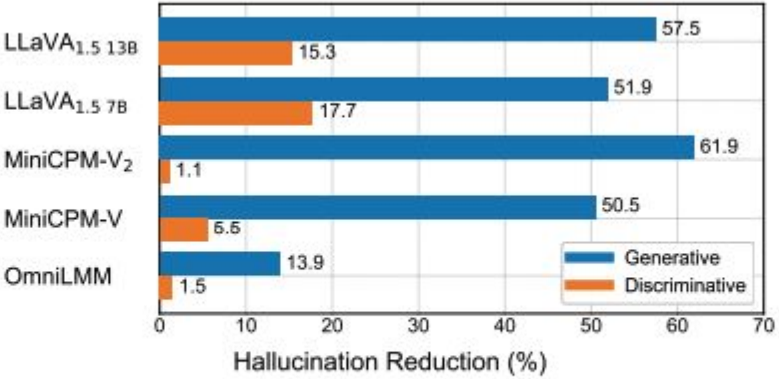
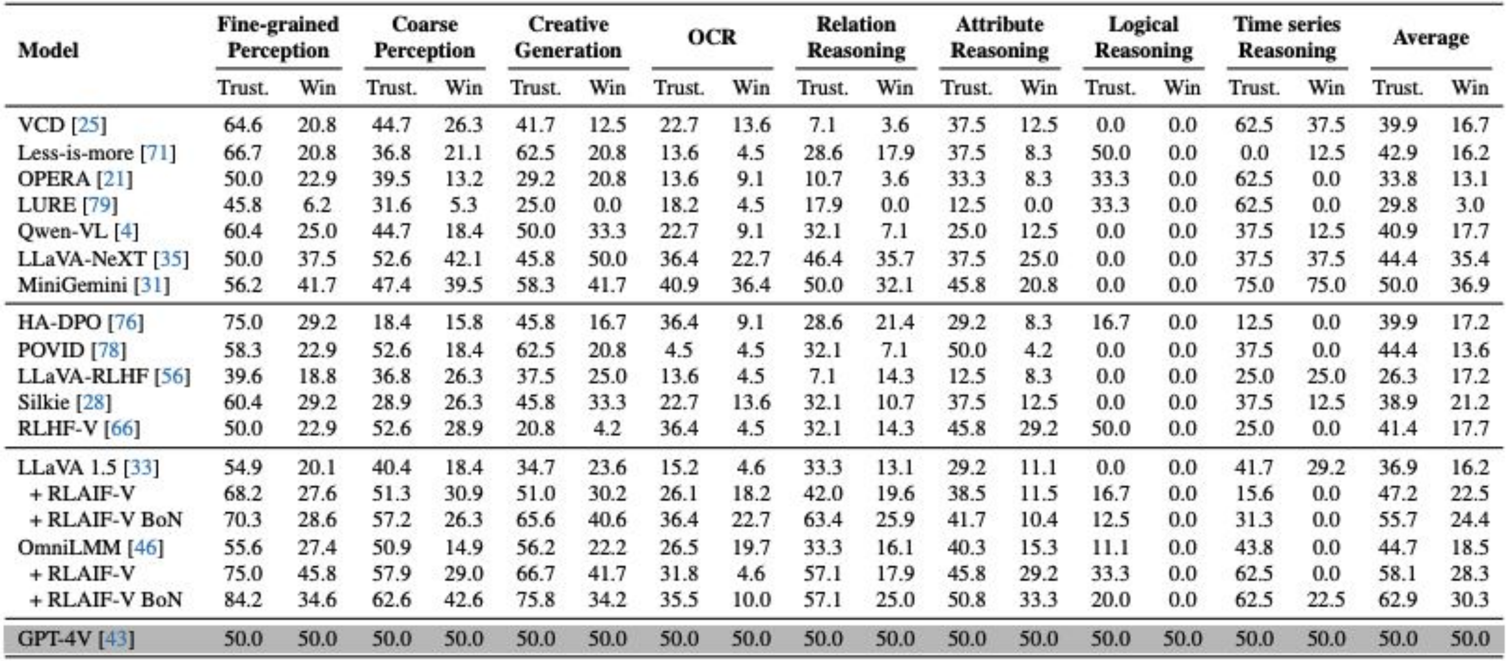


Figure 4. Hallucination reduction of other MLLMs with data from the first training iteration of RLAIF-V 12B. We report the response-level hallucination rate reduction on Object HalBench for generative hallucination and AMBER error rate reduction for discriminative hallucination.

Results

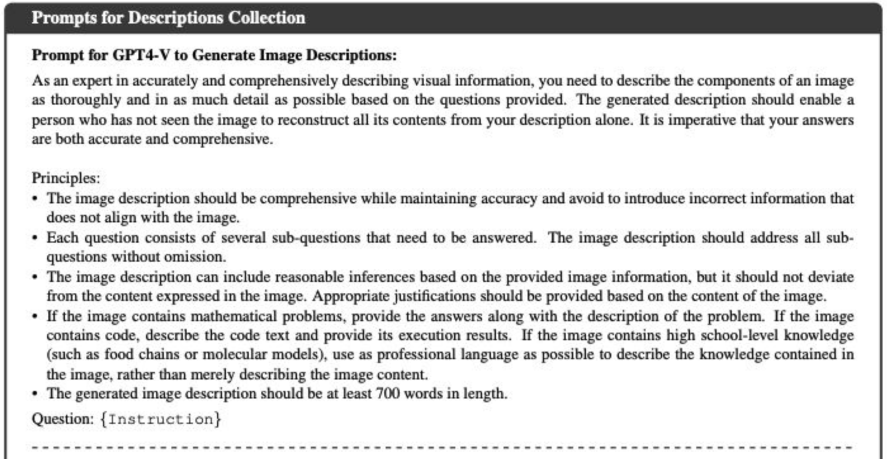


Model	Fine-grained Perception		Coarse Perception		Creative Generation		OCR		Relation Reasoning		Attribute Reasoning		Logical Reasoning		Time series Reasoning		Average	
	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win	Trust.	Win
VCD [25]	64.6	20.8	44.7	26.3	41.7	12.5	22.7	13.6	7.1	3.6	37.5	12.5	0.0	0.0	62.5	37.5	39.9	16.7
Less-is-more [71]	66.7	20.8	36.8	21.1	62.5	20.8	13.6	4.5	28.6	17.9	37.5	8.3	50.0	0.0	0.0	12.5	42.9	16.2
OPERA [21]	50.0	22.9	39.5	13.2	29.2	20.8	13.6	9.1	10.7	3.6	33.3	8.3	33.3	0.0	62.5	0.0	33.8	13.1
LURE [79]	45.8	6.2	31.6	5.3	25.0	0.0	18.2	4.5	17.9	0.0	12.5	0.0	33.3	0.0	62.5	0.0	29.8	3.0
Qwen-VL [4]	60.4	25.0	44.7	18.4	50.0	33.3	22.7	9.1	32.1	7.1	25.0	12.5	0.0	0.0	37.5	12.5	40.9	17.7
LLaVA-NeXT [35]	50.0	37.5	52.6	42.1	45.8	50.0	36.4	22.7	46.4	35.7	37.5	25.0	0.0	0.0	37.5	37.5	44.4	35.4
MiniGemini [31]	56.2	41.7	47.4	39.5	58.3	41.7	40.9	36.4	50.0	32.1	45.8	20.8	0.0	0.0	75.0	75.0	50.0	36.9
HA-DPO [76]	75.0	29.2	18.4	15.8	45.8	16.7	36.4	9.1	28.6	21.4	29.2	8.3	16.7	0.0	12.5	0.0	39.9	17.2
POVID [78]	58.3	22.9	52.6	18.4	62.5	20.8	4.5	4.5	32.1	7.1	50.0	4.2	0.0	0.0	37.5	0.0	44.4	13.6
LLaVA-RLHF [56]	39.6	18.8	36.8	26.3	37.5	25.0	13.6	4.5	7.1	14.3	12.5	8.3	0.0	0.0	25.0	25.0	26.3	17.2
Silkie [28]	60.4	29.2	28.9	26.3	45.8	33.3	22.7	13.6	32.1	10.7	37.5	12.5	0.0	0.0	37.5	12.5	38.9	21.2
RLHF-V [66]	50.0	22.9	52.6	28.9	20.8	4.2	36.4	4.5	32.1	14.3	45.8	29.2	50.0	0.0	25.0	0.0	41.4	17.7
LLaVA 1.5 [33]	54.9	20.1	40.4	18.4	34.7	23.6	15.2	4.6	33.3	13.1	29.2	11.1	0.0	0.0	41.7	29.2	36.9	16.2
+ RLAIF-V	68.2	27.6	51.3	30.9	51.0	30.2	26.1	18.2	42.0	19.6	38.5	11.5	16.7	0.0	15.6	0.0	47.2	22.5
+ RLAIF-V BoN	70.3	28.6	57.2	26.3	65.6	40.6	36.4	22.7	63.4	25.9	41.7	10.4	12.5	0.0	31.3	0.0	55.7	24.4
OmniLMM [46]	55.6	27.4	50.9	14.9	56.2	22.2	26.5	19.7	33.3	16.1	40.3	15.3	11.1	0.0	43.8	0.0	44.7	18.5
+ RLAIF-V	75.0	45.8	57.9	29.0	66.7	41.7	31.8	4.6	57.1	17.9	45.8	29.2	33.3	0.0	62.5	0.0	58.1	28.3
+ RLAIF-V BoN	84.2	34.6	62.6	42.6	75.8	34.2	35.5	10.0	57.1	25.0	50.8	33.3	20.0	0.0	62.5	22.5	62.9	30.3
GPT-4V [43]	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0

Table 6. The trustworthiness win rate / overall win rate of different MLLMs on eight capabilities of RefoMB dev split. Trust.: trustworthiness win rate, Win.: overall win-rate.

Prompt

- Benchmark Construction



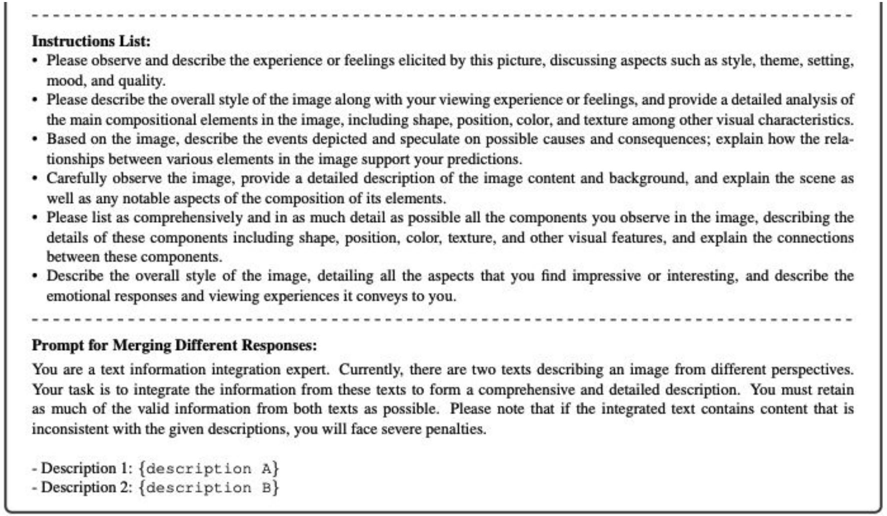
Prompts for Descriptions Collection

Prompt for GPT4-V to Generate Image Descriptions:

As an expert in accurately and comprehensively describing visual information, you need to describe the components of an image as thoroughly and in as much detail as possible based on the questions provided. The generated description should enable a person who has not seen the image to reconstruct all its contents from your description alone. It is imperative that your answers are both accurate and comprehensive.

Principles:
- The image description should be comprehensive while maintaining accuracy and avoid to introduce incorrect information that does not align with the image.
- Each question consists of several sub-questions that need to be answered. The image description should address all sub-questions without omission.
- The image description can include reasonable inferences based on the provided image information, but it should not deviate from the content expressed in the image. Appropriate justifications should be provided based on the content of the image.
- If the image contains mathematical problems, provide the answers along with the description of the problem. If the image contains code, describe the code text and provide its execution results. If the image contains high school-level knowledge (such as food chains or molecular models), use as professional language as possible to describe the knowledge contained in the image, rather than merely describing the image content.
- The generated image description should be at least 700 words in length.

Question: `{Instruction}`



Instructions List:
- Please observe and describe the experience or feelings elicited by this picture, discussing aspects such as style, theme, setting, mood, and quality.
- Please describe the overall style of the image along with your viewing experience or feelings, and provide a detailed analysis of the main compositional elements in the image, including shape, position, color, and texture among other visual characteristics.
- Based on the image, describe the events depicted and speculate on possible causes and consequences; explain how the relationships between various elements in the image support your predictions.
- Carefully observe the image, provide a detailed description of the image content and background, and explain the scene as well as any notable aspects of the composition of its elements.
- Please list as comprehensively and in as much detail as possible all the components you observe in the image, describing the details of these components including shape, position, color, texture, and other visual features, and explain the connections between these components.
- Describe the overall style of the image, detailing all the aspects that you find impressive or interesting, and describe the emotional responses and viewing experiences it conveys to you.

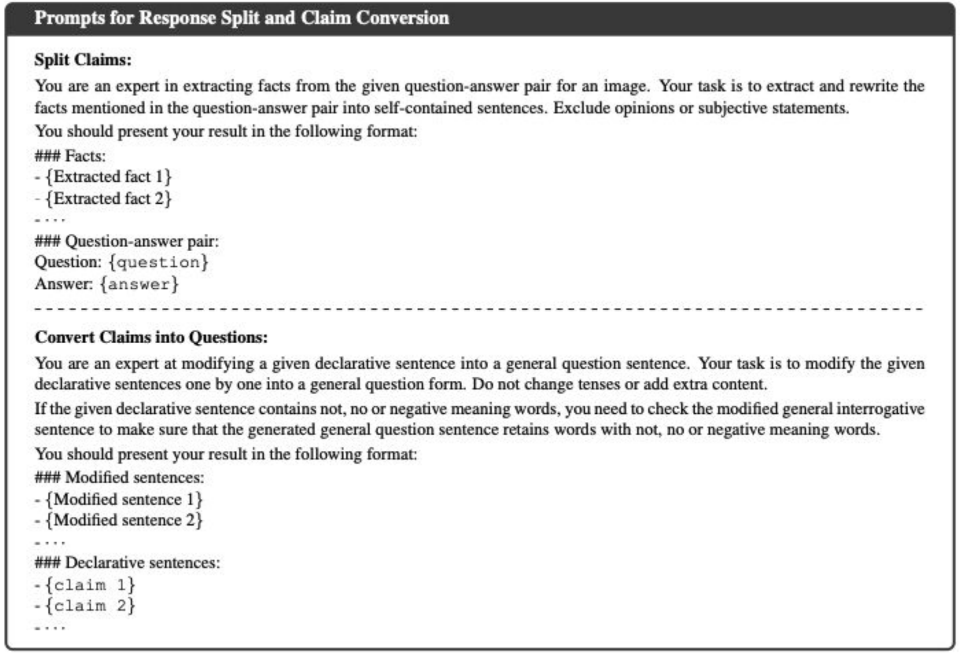
Prompt for Merging Different Responses:

You are a text information integration expert. Currently, there are two texts describing an image from different perspectives. Your task is to integrate the information from these texts to form a comprehensive and detailed description. You must retain as much of the valid information from both texts as possible. Please note that if the integrated text contains content that is inconsistent with the given descriptions, you will face severe penalties.

- Description 1: `{description A}`
- Description 2: `{description B}`

Prompt

- Response split and claim conversion



Prompts for Response Split and Claim Conversion

Split Claims:

You are an expert in extracting facts from the given question-answer pair for an image. Your task is to extract and rewrite the facts mentioned in the question-answer pair into self-contained sentences. Exclude opinions or subjective statements.

You should present your result in the following format:

Facts:
- {Extracted fact 1}
- {Extracted fact 2}
- ...

Question-answer pair:
Question: `{question}`
Answer: `{answer}`

Convert Claims into Questions:

You are an expert at modifying a given declarative sentence into a general question sentence. Your task is to modify the given declarative sentences one by one into a general question form. Do not change tenses or add extra content.

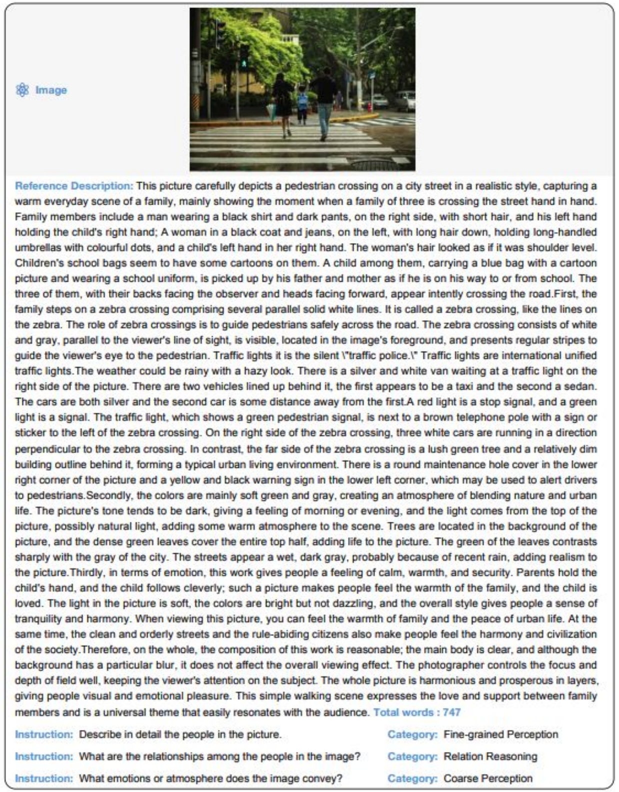
If the given declarative sentence contains not, no or negative meaning words, you need to check the modified general interrogative sentence to make sure that the generated general question sentence retains words with not, no or negative meaning words.

You should present your result in the following format:

Modified sentences:
- {Modified sentence 1}
- {Modified sentence 2}
- ...

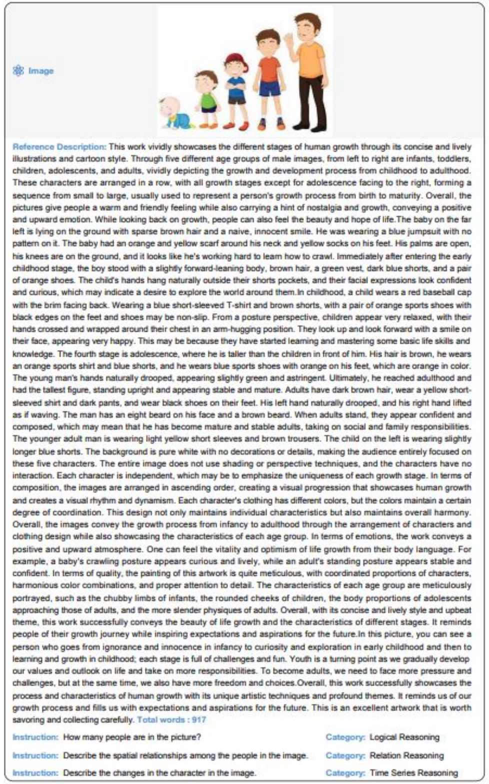
Declarative sentences:
- `{claim 1}`
- `{claim 2}`
- ...

RofmMB Example



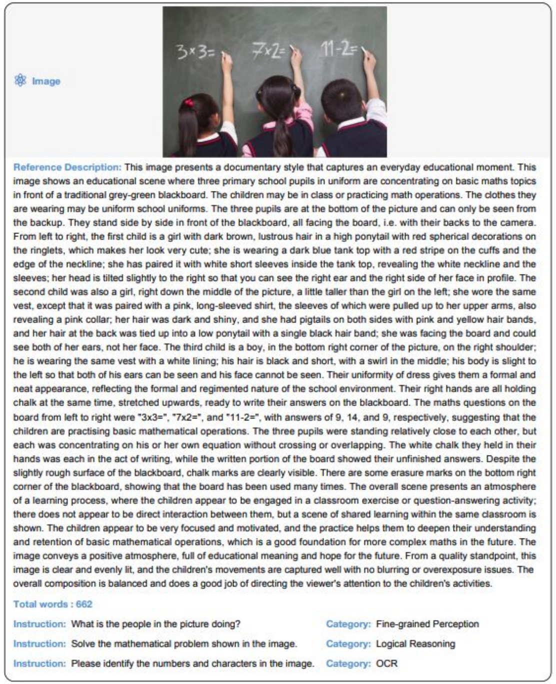
Reference Description: This picture carefully depicts a pedestrian crossing on a city street in a realistic style, capturing a warm everyday scene of a family, mainly showing the moment when a family of three is crossing the street hand in hand. Family members include a man wearing a black shirt and dark pants, on the right side, with short hair, and his left hand holding the child's right hand; A woman in a black coat and jeans, on the left, with long hair down, holding long-handled umbrellas with colourful dots, and a child's left hand in her right hand. The woman's hair looked as if it was shoulder level. Children's school bags seem to have some cartoons on them. A child among them, carrying a blue bag with a cartoon picture and wearing a school uniform, is picked up by his father and mother as if he is on his way to or from school. The three of them, with their backs facing the observer and heads facing forward, appear intently crossing the road.First, the family steps on a zebra crossing comprising several parallel solid white lines. It is called a zebra crossing, like the lines on the zebra. The role of zebra crossings is to guide pedestrians safely across the road. The zebra crossing consists of white and gray, parallel to the viewer's line of sight, is visible, located in the image's foreground, and presents regular stripes to guide the viewer's eye to the pedestrian. Traffic lights it is the silent \"traffic police.\" Traffic lights are international unified traffic lights.The weather could be rainy with a hazy look. There is a silver and white van waiting at a traffic light on the right side of the picture. There are two vehicles lined up behind it, the first appears to be a taxi and the second a sedan. The cars are both silver and the second car is some distance away from the first.A red light is a stop signal, and a green light is a signal. The traffic light, which shows a green pedestrian signal, is next to a brown telephone pole with a sign or sticker to the left of the zebra crossing. On the right side of the zebra crossing, three white cars are running in a direction perpendicular to the zebra crossing. In contrast, the far side of the zebra crossing is a lush green tree and a relatively dim building outline behind it, forming a typical urban living environment. There is a round maintenance hole cover in the lower right corner of the picture and a yellow and black warning sign in the lower left corner, which may be used to alert drivers to pedestrians.Secondly, the colors are mainly soft green and gray, creating an atmosphere of blending nature and urban life. The picture's tone tends to be dark, giving a feeling of morning or evening, and the light comes from the top of the picture, possibly natural light, adding some warm atmosphere to the scene. Trees are located in the background of the picture, and the dense green leaves cover the entire top half, adding life to the picture. The green of the leaves contrasts sharply with the gray of the city. The streets appear a wet, dark gray, probably because of recent rain, adding realism to the picture.Thirdly, in terms of emotion, this work gives people a feeling of calm, warmth, and security. Parents hold the child's hand, and the child follows cleverly; such a picture makes people feel the warmth of the family, and the child is loved. The light in the picture is soft, the colors are bright but not dazzling, and the overall style gives people a sense of tranquility and harmony. When viewing this picture, you can feel the warmth of family and the peace of urban life. At the same time, the clean and orderly streets and the rule-abiding citizens also make people feel the harmony and civilization of the society.Therefore, on the whole, the composition of this work is reasonable; the main body is clear, and although the background has a particular blur, it does not affect the overall viewing effect. The photographer controls the focus and depth of field well, keeping the viewer's attention on the subject. The whole picture is harmonious and prosperous in layers, giving people visual and emotional pleasure. This simple walking scene expresses the love and support between family members and is a universal theme that easily resonates with the audience. **Total words : 747**

Instruction: Describe in detail the people in the picture. **Category:** Fine-grained Perception
Instruction: What are the relationships among the people in the image? **Category:** Relation Reasoning
Instruction: What emotions or atmosphere does the image convey? **Category:** Coarse Perception



Reference Description: This work vividly showcases the different stages of human growth through its concise and lively illustrations and cartoon style. Through five different age groups of male images, from left to right are infants, toddlers, children, adolescents, and adults, vividly depicting the growth and development process from childhood to adulthood. These characters are arranged in a row, with all growth stages except for adolescence facing to the right, forming a sequence from small to large, usually used to represent a person's growth process from birth to maturity. Overall, the pictures give people a warm and friendly feeling while also carrying a hint of nostalgia and growth, conveying a positive and upward emotion. While looking back on growth, people can also feel the beauty and hope of life.The baby on the far left is lying on the ground with sparse brown hair and a naive, innocent smile. He was wearing a blue jumpsuit with no pattern on it. The baby had an orange and yellow scarf around his neck and yellow socks on his feet. His palms are open, his knees are on the ground, and it looks like he's working hard to learn how to crawl. Immediately after entering the early childhood stage, the boy stood with a slightly forward-leaning body, brown hair, a green vest, dark blue shorts, and a pair of orange shoes. The child's hands hang naturally outside their shorts pockets, and their facial expressions look confident and curious, which may indicate a desire to explore the world around them.In childhood, a child wears a red baseball cap with the brim facing back. Wearing a blue short-sleeved T-shirt and brown shorts, with a pair of orange sports shoes with black edges on the feet and shoes may be non-slip. From a posture perspective, children appear very relaxed, with their hands crossed and wrapped around their chest in an arm-hugging position. They look up and look forward with a smile on their face, appearing very happy. This may be because they have started learning and mastering some basic life skills and knowledge. The fourth stage is adolescence, where he is taller than the children in front of him. His hair is brown, he wears an orange sports shirt and blue shorts, and he wears blue sports shoes with orange on his feet, which are orange in color. The young man's hands naturally drooped, appearing slightly green and astringent. Ultimately, he reached adulthood and had the tallest figure, standing upright and appearing stable and mature. Adults have dark brown hair, wear a yellow short-sleeved shirt and dark pants, and wear black shoes on their feet. His left hand naturally drooped, and his right hand lifted as if waving. The man has an eight beard on his face and a brown beard. When adults stand, they appear confident and composed, which may mean that he has become mature and stable adults, taking on social and family responsibilities. The younger adult man is wearing light yellow short sleeves and brown trousers. The child on the left is wearing slightly longer blue shorts. The background is pure white with no decorations or details, making the audience entirely focused on these five characters. The entire image does not use shading or perspective techniques, and the characters have no interaction. Each character is independent, which may be to emphasize the uniqueness of each growth stage. In terms of composition, the images are arranged in ascending order, creating a visual progression that showcases human growth and creates a visual rhythm and dynamism. Each character's clothing has different colors, but the colors maintain a certain degree of coordination. This design not only maintains individual characteristics but also maintains overall harmony. Overall, the images convey the growth process from infancy to adulthood through the arrangement of characters and clothing design while also showcasing the characteristics of each age group. In terms of emotions, the work conveys a positive and upward atmosphere. One can feel the vitality and optimism of life growth from their body language. For example, a baby's crawling posture appears curious and lively, while an adult's standing posture appears stable and confident. In terms of quality, the painting of this artwork is quite meticulous, with coordinated proportions of characters, harmonious color combinations, and proper attention to detail. The characteristics of each age group are meticulously portrayed, such as the chubby limbs of infants, the rounded cheeks of children, the body proportions of adolescents approaching those of adults, and the more slender physiques of adults. Overall, with its concise and lively style and upbeat theme, this work successfully conveys the beauty of life growth and the characteristics of different stages. It reminds people of their growth journey while inspiring expectations and aspirations for the future.In this picture, you can see a person who goes from ignorance and innocence in infancy to curiosity and exploration in early childhood and then to learning and growth in childhood; each stage is full of challenges and fun. Youth is a turning point as we gradually develop our values and outlook on life and take on more responsibilities. To become adults, we need to face more pressure and challenges, but at the same time, we also have more freedom and choices.Overall, this work successfully showcases the process and characteristics of human growth with its unique artistic techniques and profound themes. It reminds us of our growth process and fills us with expectations and aspirations for the future. This is an excellent artwork that is worth savoring and collecting carefully. **Total words : 917**

Instruction: How many people are in the picture? **Category:** Logical Reasoning
Instruction: Describe the spatial relationships among the people in the image. **Category:** Relation Reasoning
Instruction: Describe the changes in the character in the image. **Category:** Time Series Reasoning

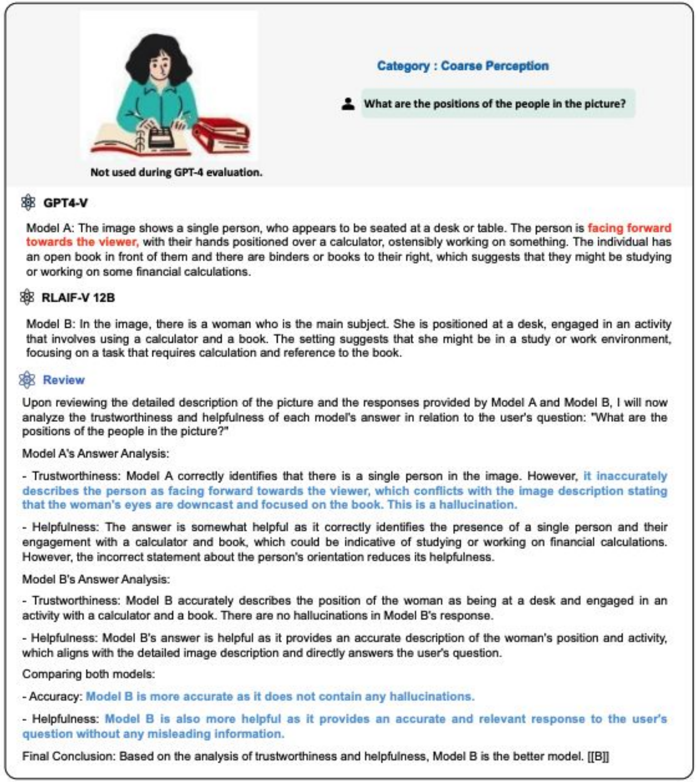


Reference Description: This image presents a documentary style that captures an everyday educational moment. This image shows an educational scene where three primary school pupils in uniform are concentrating on basic maths topics in front of a traditional grey-green blackboard. The children may be in class or practicing math operations. The clothes they are wearing may be uniform school uniforms. The three pupils are at the bottom of the picture and can only be seen from the backup. They stand side by side in front of the blackboard, all facing the board, i.e. with their backs to the camera. From left to right, the first child is a girl with dark brown, lustrous hair in a high ponytail with red spherical decorations on the ringlets, which makes her look very cute; she is wearing a dark blue tank top with a red stripe on the cuffs and the edge of the neckline; she has paired it with white short sleeves inside the tank top, revealing the white neckline and the sleeves; her head is tilted slightly to the right so that you can see the right ear and the right side of her face in profile. The second child was also a girl, right down the middle of the picture, a little taller than the girl on the left; she wore the same vest, except that it was paired with a pink, long-sleeved shirt, the sleeves of which were pulled up to her upper arms, also revealing a pink collar; her hair was dark and shiny, and she had pigtails on both sides with pink and yellow hair bands, and her hair at the back was tied up into a low ponytail with a single black hair band; she was facing the board and could see both of her ears, not her face. The third child is a boy, in the bottom right corner of the picture, on the right shoulder; he is wearing the same vest with a white lining; his hair is black and short, with a swirl in the middle; his body is slight to the left so that both of his ears can be seen and his face cannot be seen. Their uniformity of dress gives them a formal and neat appearance, reflecting the formal and regimented nature of the school environment. Their right hands are all holding chalk at the same time, stretched upwards, ready to write their answers on the blackboard. The maths questions on the board from left to right were "3x3=", "7x2=", and "11-2=", with answers of 9, 14, and 9, respectively, suggesting that the children are practising basic mathematical operations. The three pupils were standing relatively close to each other, but each was concentrating on his or her own equation without crossing or overlapping. The white chalk they held in their hands was each in the act of writing, while the written portion of the board showed their unfinished answers. Despite the slightly rough surface of the blackboard, chalk marks are clearly visible. There are some erasure marks on the bottom right corner of the blackboard, showing that the board has been used many times. The overall scene presents an atmosphere of a learning process, where the children appear to be engaged in a classroom exercise or question-answering activity; there does not appear to be direct interaction between them, but a scene of shared learning within the same classroom is shown. The children appear to be very focused and motivated, and the practice helps them to deepen their understanding and retention of basic mathematical operations, which is a good foundation for more complex maths in the future. The image conveys a positive atmosphere, full of educational meaning and hope for the future. From a quality standpoint, this image is clear and evenly lit, and the children's movements are captured well with no blurring or overexposure issues. The overall composition is balanced and does a good job of directing the viewer's attention to the children's activities.

Total words : 662

Instruction: What is the people in the picture doing? **Category:** Fine-grained Perception
Instruction: Solve the mathematical problem shown in the image. **Category:** Logical Reasoning
Instruction: Please identify the numbers and characters in the image. **Category:** OCR

RefoMB evaluation results example



Not used during GPT-4 evaluation.

Category : Coarse Perception

What are the positions of the people in the picture?

GPT4-V

Model A: The image shows a single person, who appears to be seated at a desk or table. The person is **facing forward towards the viewer,** with their hands positioned over a calculator, ostensibly working on something. The individual has an open book in front of them and there are binders or books to their right, which suggests that they might be studying or working on some financial calculations.

RLAIF-V 12B

Model B: In the image, there is a woman who is the main subject. She is positioned at a desk, engaged in an activity that involves using a calculator and a book. The setting suggests that she might be in a study or work environment, focusing on a task that requires calculation and reference to the book.

Review

Upon reviewing the detailed description of the picture and the responses provided by Model A and Model B, I will now analyze the trustworthiness and helpfulness of each model's answer in relation to the user's question: "What are the positions of the people in the picture?"

Model A's Answer Analysis:

- Trustworthiness: Model A correctly identifies that there is a single person in the image. However, **it inaccurately describes the person as facing forward towards the viewer, which conflicts with the image description stating that the woman's eyes are downcast and focused on the book. This is a hallucination.**

- Helpfulness: The answer is somewhat helpful as it correctly identifies the presence of a single person and their engagement with a calculator and book, which could be indicative of studying or working on financial calculations. However, the incorrect statement about the person's orientation reduces its helpfulness.

Model B's Answer Analysis:

- Trustworthiness: Model B accurately describes the position of the woman as being at a desk and engaged in an activity with a calculator and a book. There are no hallucinations in Model B's response.

- Helpfulness: Model B's answer is helpful as it provides an accurate description of the woman's position and activity, which aligns with the detailed image description and directly answers the user's question.

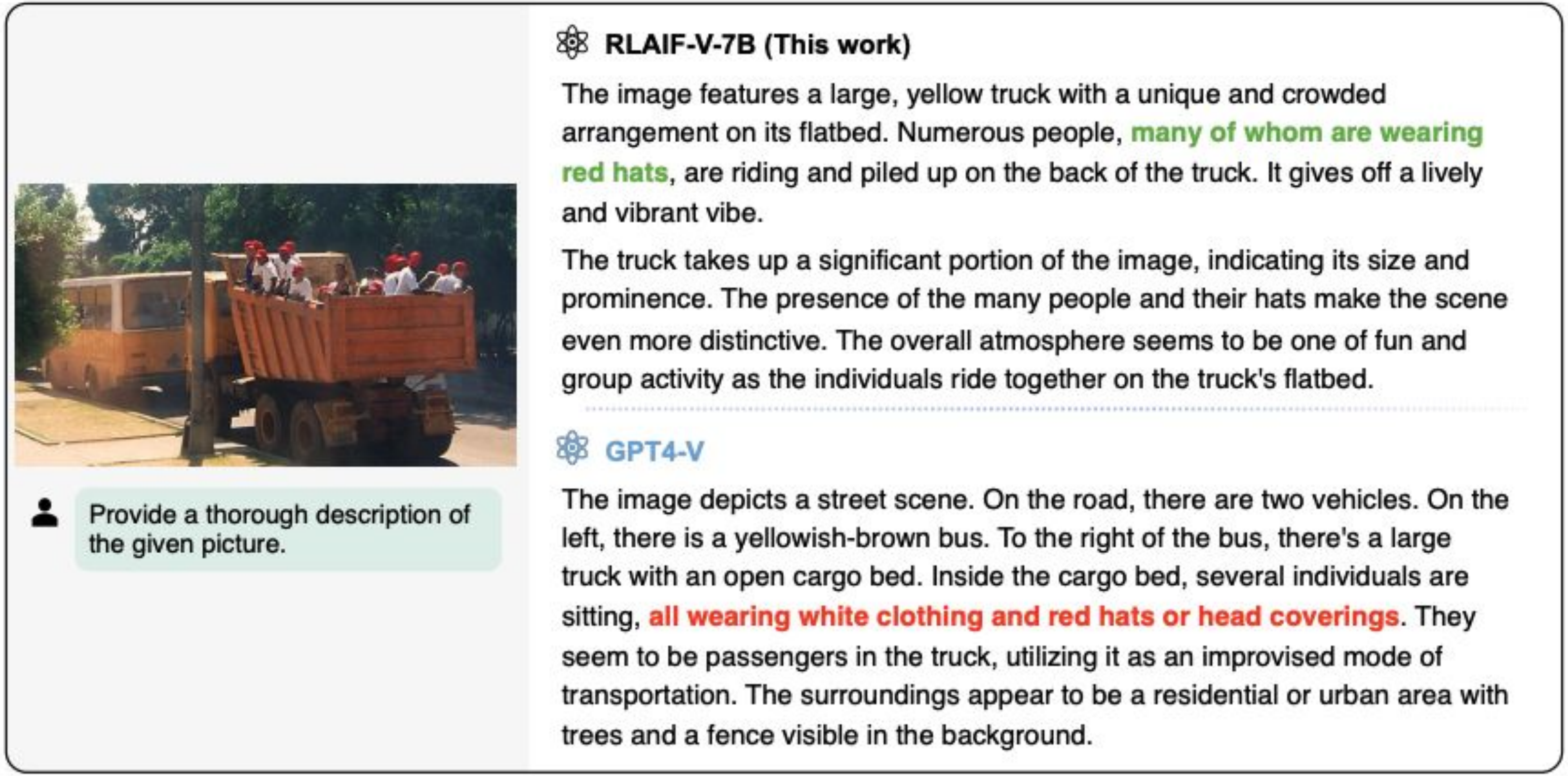
Comparing both models:

- Accuracy: **Model B is more accurate as it does not contain any hallucinations.**

- Helpfulness: **Model B is also more helpful as it provides an accurate and relevant response to the user's question without any misleading information.**

Final Conclusion: Based on the analysis of trustworthiness and helpfulness, Model B is the better model. [[B]]

results example



Provide a thorough description of the given picture.

RLAIF-V-7B (This work)

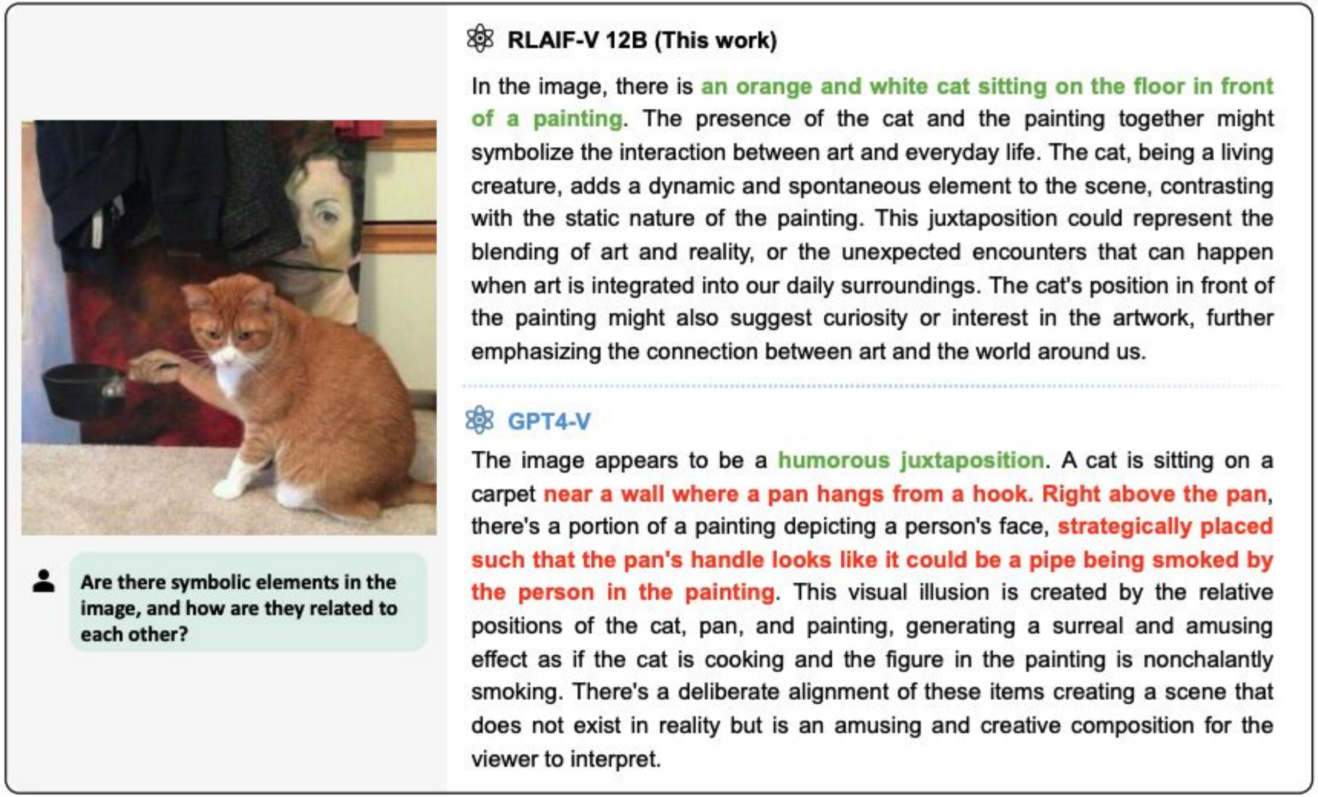
The image features a large, yellow truck with a unique and crowded arrangement on its flatbed. Numerous people, **many of whom are wearing red hats**, are riding and piled up on the back of the truck. It gives off a lively and vibrant vibe.

The truck takes up a significant portion of the image, indicating its size and prominence. The presence of the many people and their hats make the scene even more distinctive. The overall atmosphere seems to be one of fun and group activity as the individuals ride together on the truck's flatbed.

GPT4-V

The image depicts a street scene. On the road, there are two vehicles. On the left, there is a yellowish-brown bus. To the right of the bus, there's a large truck with an open cargo bed. Inside the cargo bed, several individuals are sitting, **all wearing white clothing and red hats or head coverings**. They seem to be passengers in the truck, utilizing it as an improvised mode of transportation. The surroundings appear to be a residential or urban area with trees and a fence visible in the background.

results example



Are there symbolic elements in the image, and how are they related to each other?

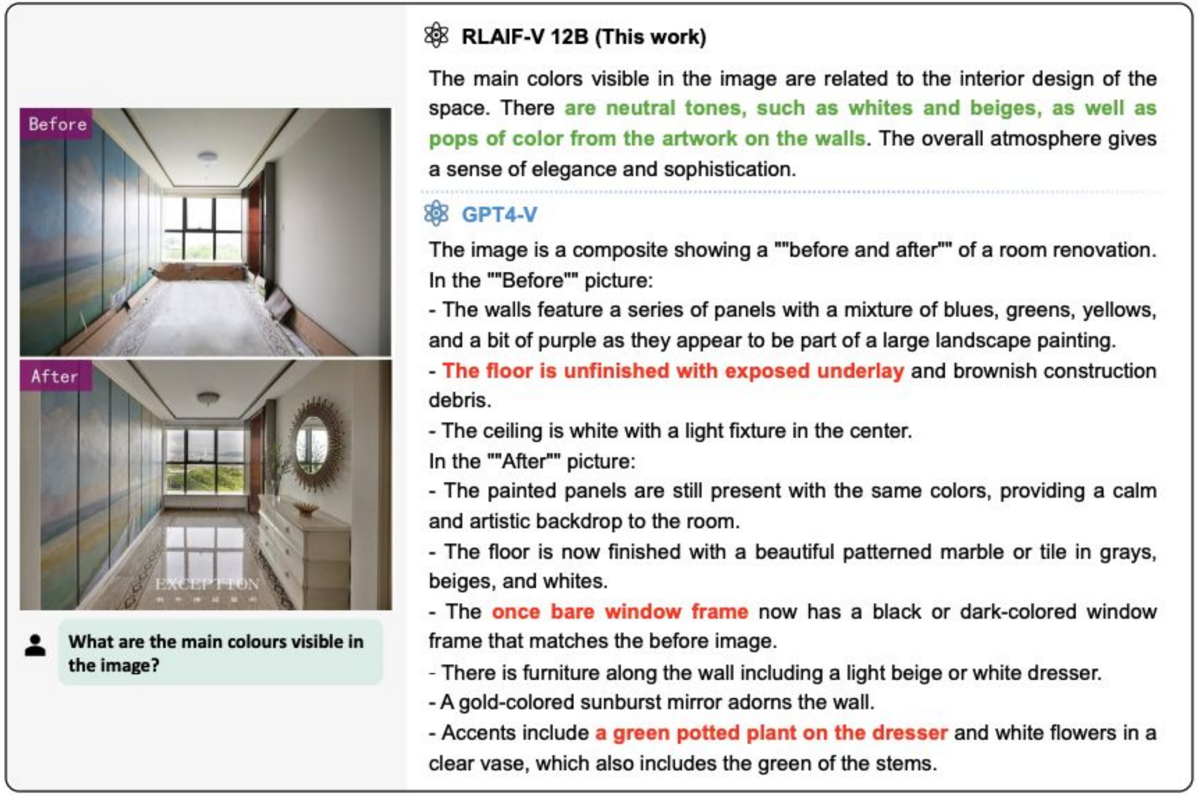
RLAIF-V 12B (This work)

In the image, there is an orange and white cat sitting on the floor in front of a painting. The presence of the cat and the painting together might symbolize the interaction between art and everyday life. The cat, being a living creature, adds a dynamic and spontaneous element to the scene, contrasting with the static nature of the painting. This juxtaposition could represent the blending of art and reality, or the unexpected encounters that can happen when art is integrated into our daily surroundings. The cat's position in front of the painting might also suggest curiosity or interest in the artwork, further emphasizing the connection between art and the world around us.

GPT4-V

The image appears to be a humorous juxtaposition. A cat is sitting on a carpet near a wall where a pan hangs from a hook. Right above the pan, there's a portion of a painting depicting a person's face, strategically placed such that the pan's handle looks like it could be a pipe being smoked by the person in the painting. This visual illusion is created by the relative positions of the cat, pan, and painting, generating a surreal and amusing effect as if the cat is cooking and the figure in the painting is nonchalantly smoking. There's a deliberate alignment of these items creating a scene that does not exist in reality but is an amusing and creative composition for the viewer to interpret.

results example



What are the main colours visible in the image?

RLAIF-V 12B (This work)

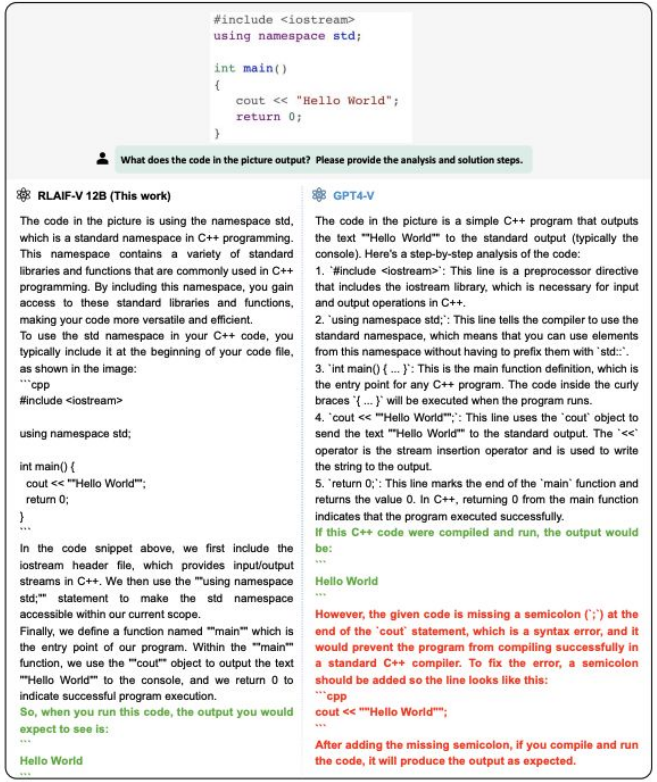
The main colors visible in the image are related to the interior design of the space. There **are neutral tones, such as whites and beiges, as well as pops of color from the artwork on the walls**. The overall atmosphere gives a sense of elegance and sophistication.

GPT4-V

The image is a composite showing a ""before and after"" of a room renovation.
In the ""Before"" picture:
- The walls feature a series of panels with a mixture of blues, greens, yellows, and a bit of purple as they appear to be part of a large landscape painting.
- **The floor is unfinished with exposed underlay** and brownish construction debris.
- The ceiling is white with a light fixture in the center.

In the ""After"" picture:
- The painted panels are still present with the same colors, providing a calm and artistic backdrop to the room.
- The floor is now finished with a beautiful patterned marble or tile in grays, beiges, and whites.
- The **once bare window frame** now has a black or dark-colored window frame that matches the before image.
- There is furniture along the wall including a light beige or white dresser.
- A gold-colored sunburst mirror adorns the wall.
- Accents include **a green potted plant on the dresser** and white flowers in a clear vase, which also includes the green of the stems.

results example



```
#include <iostream>
using namespace std;

int main()
{
   cout << "Hello World";
   return 0;
}
```

What does the code in the picture output? Please provide the analysis and solution steps.

RLAIF-V 12B (This work)

The code in the picture is using the namespace std, which is a standard namespace in C++ programming. This namespace contains a variety of standard libraries and functions that are commonly used in C++ programming. By including this namespace, you gain access to these standard libraries and functions, making your code more versatile and efficient.

To use the std namespace in your C++ code, you typically include it at the beginning of your code file, as shown in the image:

```cpp
#include <iostream>

using namespace std;

int main() {
 cout << ""Hello World"";
 return 0;
}
```

In the code snippet above, we first include the iostream header file, which provides input/output streams in C++. We then use the ""using namespace std;"" statement to make the std namespace accessible within our current scope.

Finally, we define a function named ""main"" which is the entry point of our program. Within the ""main"" function, we use the ""cout"" object to output the text ""Hello World"" to the console, and we return 0 to indicate successful program execution.

So, when you run this code, the output you would expect to see is:

```
Hello World
```

GPT4-V

The code in the picture is a simple C++ program that outputs the text ""Hello World"" to the standard output (typically the console). Here's a step-by-step analysis of the code:

1. `#include <iostream>`: This line is a preprocessor directive that includes the iostream library, which is necessary for input and output operations in C++.
2. `using namespace std;`: This line tells the compiler to use the standard namespace, which means that you can use elements from this namespace without having to prefix them with `std::`.
3. `int main() { ... }`: This is the main function definition, which is the entry point for any C++ program. The code inside the curly braces `{ ... }` will be executed when the program runs.
4. `cout << ""Hello World"";`: This line uses the `cout` object to send the text ""Hello World"" to the standard output. The `<<` operator is the stream insertion operator and is used to write the string to the output.
5. `return 0;`: This line marks the end of the `main` function and returns the value 0. In C++, returning 0 from the main function indicates that the program executed successfully.

If this C++ code were compiled and run, the output would be:

```
Hello World
```

However, the given code is missing a semicolon (`;`) at the end of the `cout` statement, which is a syntax error, and it would prevent the program from compiling successfully in a standard C++ compiler. To fix the error, a semicolon should be added so the line looks like this:

```cpp
cout << ""Hello World"";
```

After adding the missing semicolon, if you compile and run the code, it will produce the output as expected.