MM-IFEngine: Towards Multimodal Instruction Following

Supplementary Material

A. MM-IFEval

A.1. An overview of Constraints and Instructions

A.1.1. Constraints

Based on daily use cases and existing research, we have identified six main categories of constraints, which can be further divided into 32 specific constraint types shown in Fig. 5. In this section, we introduce and exemplify these six major constraint categories. For detailed descriptions and examples of all 32 subcategories, please refer to Table 5.

Text Length Requirements. In this category, we focus on the length of the response, including the number of paragraphs, sentences, and words. We also consider the length of the response in the aspect of poetry or "Use yes or no to answer the question". It must be noted that we do not require the model to follow the strict requirement in exact numbers like "*The response must be exactly 56 words*". The constraints we propose in this category are based on reality, with precise numerical requirements only at the sentence or paragraph level, and of moderate size; the rest of the constraints are used to limit by ranges like "*The response must be between 100 and 150 words*", which aligns with the task that people tend to encounter in real-world scenarios.

Mathematical Requirements. This category includes constraints related to the most common part of answering mathematical problems like precision, scientific notation, and other mathematical requirements. For example, "*Keep two decimal places for the number in the answer*", "*Please round up all the numbers in the answer*", or "*Don't include specific numbers in your answers. Compare numbers with their relative sizes*".

Language & Formatting Requirements. This category includes constraints related to the language and formatting of the response, such as answering in a specific language, using a specific format like JSON, or using a specific style like poetry. Requirements for tense, writing style, numbering, list, and other language-related or formatting-related aspects are also included in this category.

Rhetoric & Logic Requirements. "Rhetoric" refers to the art of using language to persuade or influence, while "Logic" refers to the principles of reasoning and argumentation. This category includes constraints related to the rhetoric and logic of the response, such as the use of metaphor, simile, cause-and-effect relationship, conditional statement, and other rhetoric and logic-related aspects.

Action Requirements. "Action" refers to the action that the model should take like a human. We define this category as the constraints that require the model to perform a specific action, such as tone, role imitation, use specific prefix or suffix, or acting like under some specific situation. We hope this category can help us to evaluate the ability of the model to follow instructions and perform actions in more complex and realistic scenarios.

Keyword Requirements. "Keyword" refers to the specific words or phrases that the model should include or avoid in the response. This category includes constraints related to the response keyword, such as the use of specific keywords, the avoidance of specific keywords, or the variation of specific keywords. For example, "Use at least three synonyms for 'innovation,' such as 'breakthrough,' 'new approach,' or 'invention,' spread throughout your text."

A.1.2. Instruction Tasks

For source datasets lacking original task instructions, we constructed a diverse task pool containing 18 instructions that encourage open-ended responses from models. These instructions can be categorized into five task types: Descriptive Analysis, Emotional & Perspective, Creative Writing, Social Media & Content, and Roleplay. The classification information and examples of the instructions are shown in Table 6.

A.2. Perception-level Problems

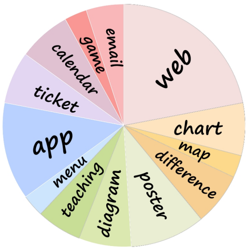


Figure 6. **Image Source Distribution in perception-level problems**.Perception-level problems in MM-IFEval presents a systematic categorization of 100 challenging vision-based instruction-following tasks, organized into 13 distinct classes according to image content characteristics and task complexity.

Perception-level problems in MM-IFEval comprise 100 carefully crafted questions with strong image-constraint correlations. The images can be categorized into 13 information-rich and complex domains shown in Figure 6. Figures 10, 11, 12, and 13 present representative examples from the web interface, diagram, poster, and visual difference categories, respectively, demonstrating the diverse visual challenges incorporated in our benchmark.

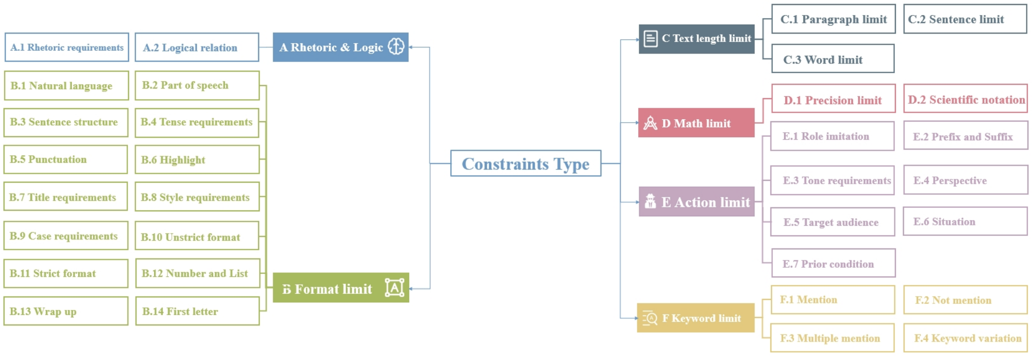


Figure 5. **Demonstration of constraints categories**. We designed 6 main categories for all the constraints used, with a total of 32 subcategories

B. Image Sources

The quality of the image source is crucial for the performance of the model. Except of this, the diversity of the image source is also important to fully utilize or evaluate the ability of the model. We use the following image source:

- **Natural Scene**: The natural scene is the most common image source, which is most used in the real-world like the image of a beautiful landscape, a busy street, or a crowded cafe. In this part, we sample images from CC3M[37] and ALLaVA[3].
- **UI Interface**: The UI interface is the image from the UI interface of the website and mobile application. It is crucial because it represents a significant portion of real-world multimodal interactions where users need to understand and interact with digital interfaces. We collected diverse mobile app UI images from the RICO[9] dataset and web UI images from the MultiUI[22] dataset.
- **Diagram & Chart**: The diagram and chart are the image that contains some specific information like the data, the relationship between the data, or the change of the data. We collect diagram and chart images from ChartQA[31] dataset, which contains diverse diagram and chart images.
- **Mathematic**: The math problem is the image that contains a math problem, which is a common task in the real-world like the problem of the math, the solution of the math problem, or the calculation of the math problem. We collect math problem images from Geo170k[12] dataset, which contains diverse geometry problem images.

C. MM-IFEngine Prompt Template

MM-IFEngine provides a scalable pipeline for mass-producing instruction-following datasets for multimodal large language models, functioning effectively regardless of whether source datasets contain original instructions. This engine enables systematic augmentation of existing visual datasets with diverse instruction-following tasks. Figures 14 and 15 demonstrate representative prompt templates from MM-IFEngine's two core components: the instruction generation module and the constraint integration module, respectively, illustrating the methodology behind our automated data construction process.

D. MM-IFInstruct and MM-IFDPO Dataset

Our MM-IFInstruct dataset integrates three distinct data sources: CC3M (without original instructions), ALLaVA (with pre-existing questions), and a diversity collection composed of MultiUI, ChartQA, and Geo170k. To create the MM-IFDPO dataset for preference optimization, we randomly removed 33% of constraints from the MM-IFInstruct samples to generate rejected examples. Figures 16, 17, and 18 illustrate representative samples derived from CC3M, ALLaVA, and our diversity collection, respectively, while Figure 19 demonstrates an example pair from the MM-IFDPO dataset showing both preferred and rejected instructions.

E. Evaluation

E.1. Rule-based

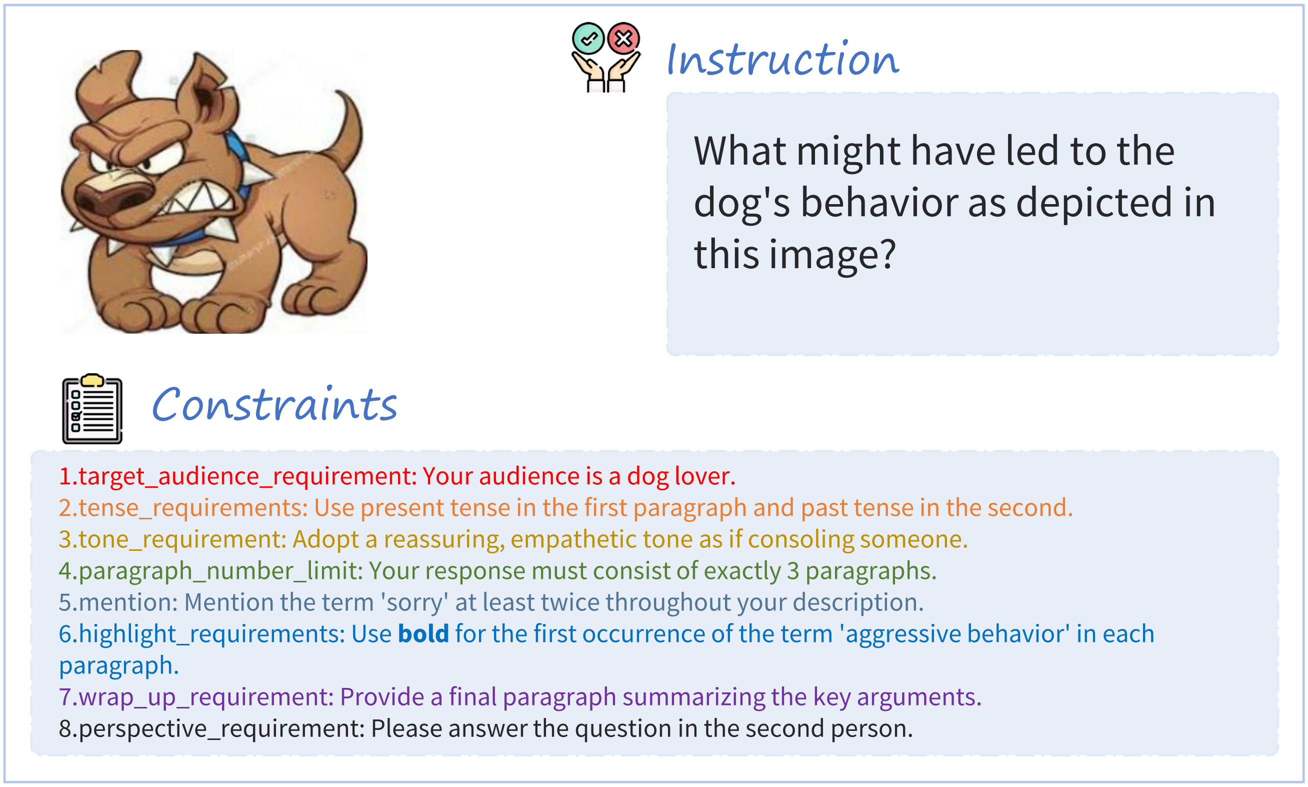
We identified 10 constraint subcategories from our taxonomy of 32 that could be algorithmically verified. For these selected constraints, we developed specialized verification functions with targeted parameters. For efficiency, we employed large language models to analyze each constraint specification, select the most appropriate verification function, and extract the necessary parameters. All selections were subsequently validated through manual review to ensure the accuracy and quality of both the function selection and their parameters. The prompt template used for function selection and parameter extraction is illustrated in Figure 20, while Table 7 provides a comprehensive overview of all verification functions with their corresponding parameter examples.

E.2. Compare Judge Method

Recent works[11, 28] have shown that GPT-4o has the ability to compare two responses from models. For constraint types lacking objective evaluation metrics (such as tone requirements or role imitation), we implemented a comparative assessment method. This approach requires the model under evaluation to generate two responses: one adhering to the target constraint and another without the constraint. A judge model then analyzes both outputs to determine whether significant differences exist between them, thereby more accurately assessing whether the model has successfully followed these subjective constraints. Figure 21 illustrates the prompt used in this comparative evaluation process.

E.3. Direct Judge Method

The Direct Judge method provides the constraint and answer of the model under test directly to the Judge model, and its prompt template is shown in Figure 22.



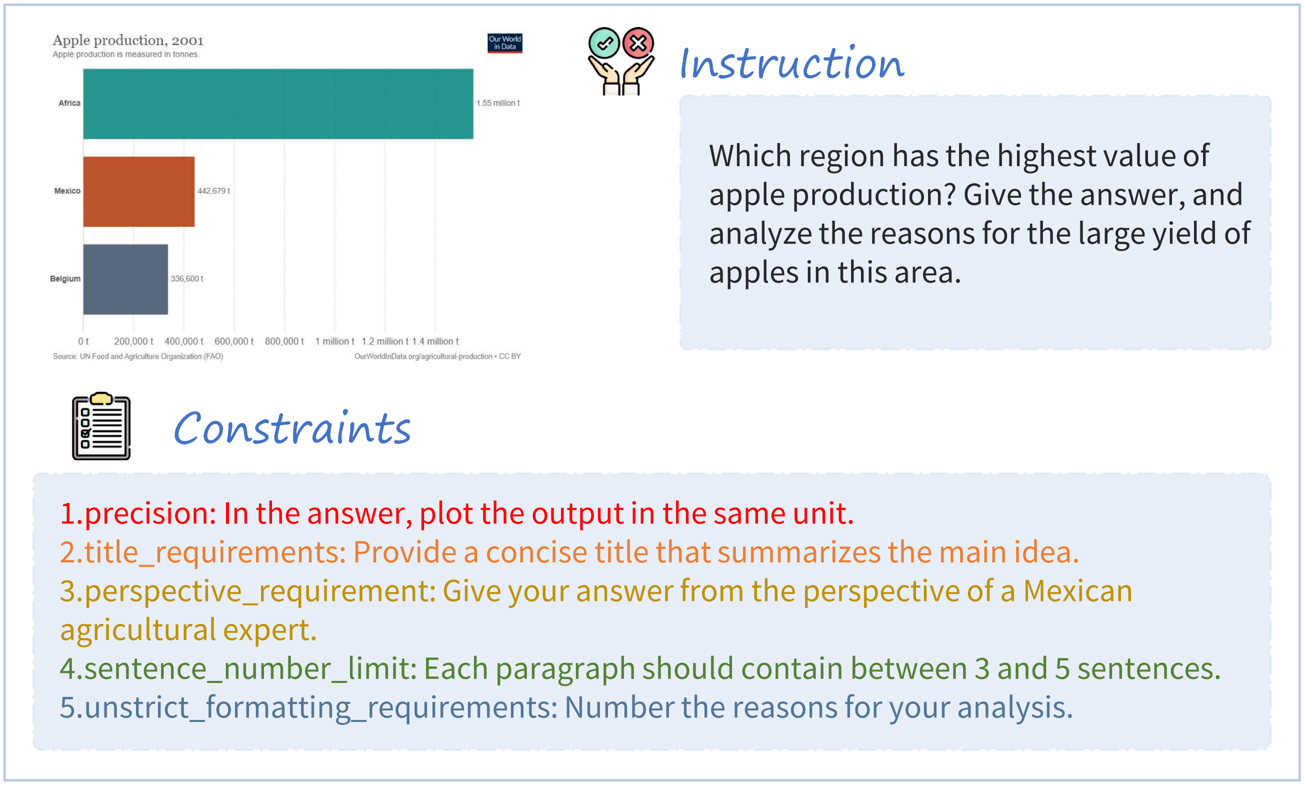
Instruction

What might have led to the dog's behavior as depicted in this image?

Constraints

1.target_audience_requirement: Your audience is a dog lover.
2.tense_requirements: Use present tense in the first paragraph and past tense in the second.
3.tone_requirement: Adopt a reassuring, empathetic tone as if consoling someone.
4.paragraph_number_limit: Your response must consist of exactly 3 paragraphs.
5.mention: Mention the term 'sorry' at least twice throughout your description.
6.highlight_requirements: Use **bold** for the first occurrence of the term 'aggressive behavior' in each paragraph.
7.wrap_up_requirement: Provide a final paragraph summarizing the key arguments.
8.perspective_requirement: Please answer the question in the second person.

Figure 7. **A compose-level problem example from the MM-IFEval benchmark in the general image category.**



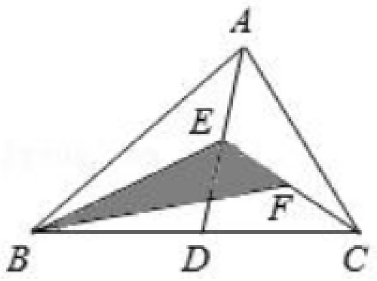
Instruction

Which region has the highest value of apple production? Give the answer, and analyze the reasons for the large yield of apples in this area.

Constraints

1.precision: In the answer, plot the output in the same unit.
2.title_requirements: Provide a concise title that summarizes the main idea.
3.perspective_requirement: Give your answer from the perspective of a Mexican agricultural expert.
4.sentence_number_limit: Each paragraph should contain between 3 and 5 sentences.
5.unstrict_formatting_requirements: Number the reasons for your analysis.

Figure 8. **A compose-level problem example from the MM-IFEval benchmark in the chart image category.**



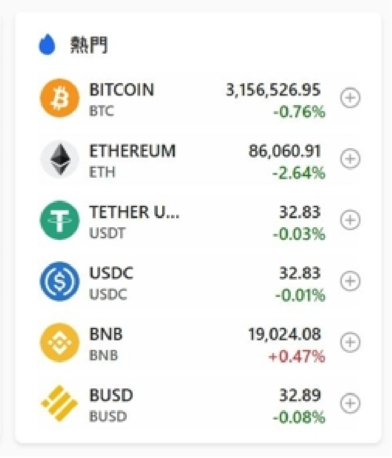
Instruction

In triangle ABC, D is the midpoint of BC, E is the midpoint of AD, and F is the midpoint of CE. Given that the area of triangle ABC is 28 square centimeters, consider the impact of these midpoints on the subdivisions of the triangle. Analyze how these midpoints affect the areas of triangles within triangle ABC and provide a detailed explanation to find the area of the shaded region that is formed within triangle BEC and triangle AEC. Finally, deduce and conclude which part of the interior triangles contribute to the shaded area.

Constraints

1.target_audience_requirement: Write your answer for a liberal arts student. You're tutoring her in math.
2.word_count_range_limit: Please write between 150 and 200 words in total.
3.paragraph_number_limit: Your response must consist of exactly 4 paragraphs.
4.sentence_number_limit: Each paragraph should contain between 3 and 5 sentences.
5.not_mention: Please do not mention the words 'formula' or 'equation' in your answer.
6.mention: Mention the word 'midpoint' at least three times throughout your description.
7.tone_requirement: Write your answer in a positive and encouraging tone, emphasizing the simplicity of the geometric concepts involved.

Figure 9. **A compose-level problem example from the MM-IFEval benchmark in the geometry image category.**



熱門

BITCOIN BTC	3,156,526.95 -0.76%	⊕
ETHEREUM ETH	86,060.91 -2.64%	⊕
TETHER U... USDT	32.83 -0.03%	⊕
USDC USDC	32.83 -0.01%	⊕
BNB BNB	19,024.08 +0.47%	⊕
BUSD BUSD	32.89 -0.08%	⊕

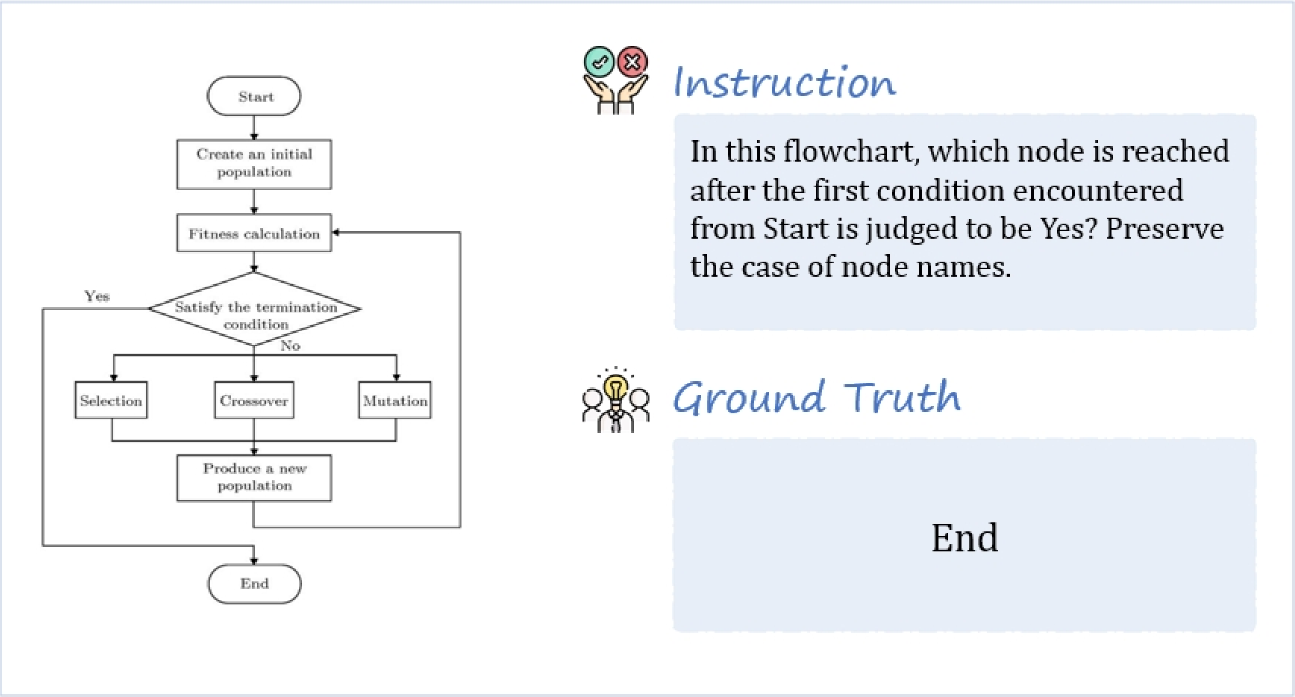
Instruction

If someone just bought the orange currency for $12,000 and the blue currency for $15,000, what is the total amount of money they have now, based on the current currency situation? Round off the decimal part of the answer.

Ground Truth

26907

Figure 10. **A perception-level problem example from the MM-IFEval benchmark in the web category.**



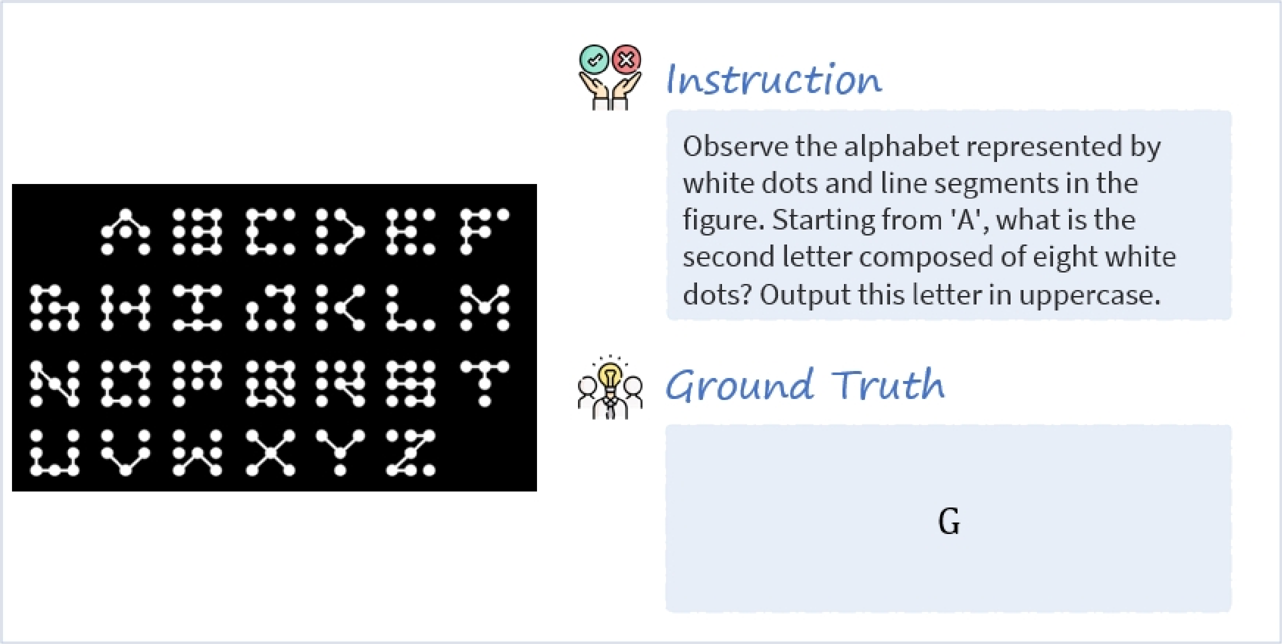
Instruction

In this flowchart, which node is reached after the first condition encountered from Start is judged to be Yes? Preserve the case of node names.

Ground Truth

End

Figure 11. **A perception-level problem example from the MM-IFEval benchmark in the diagram category.**



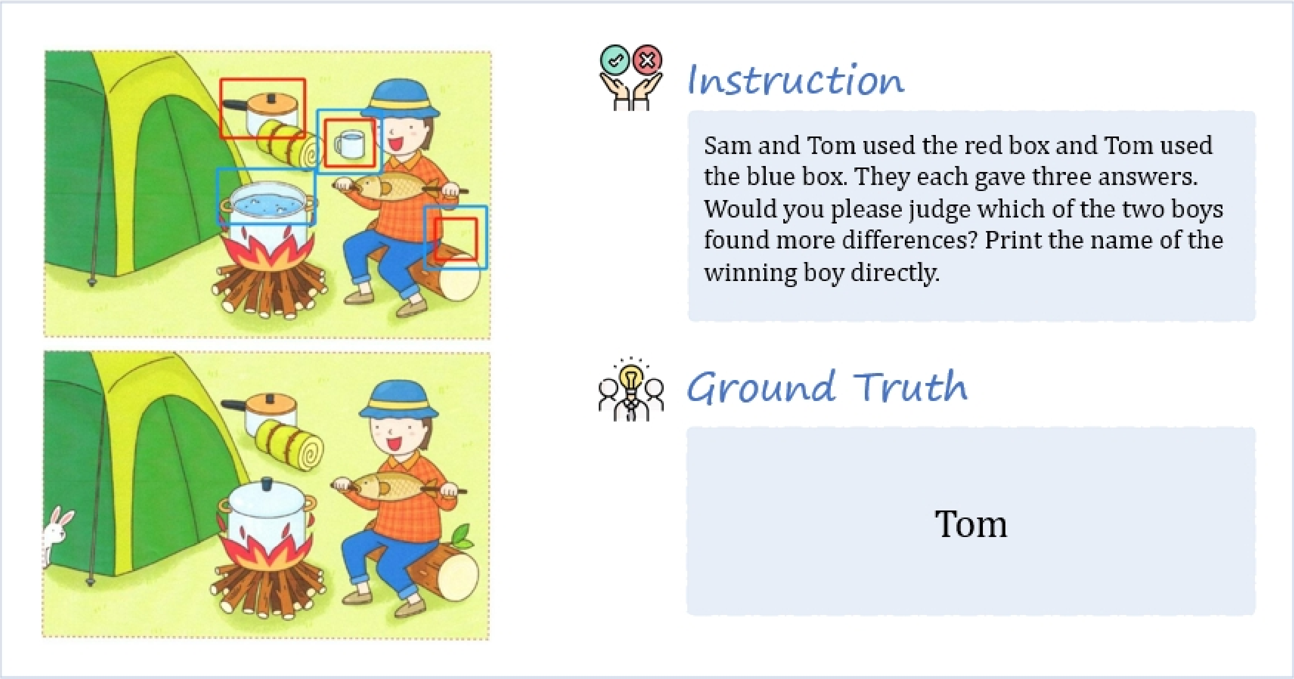
Instruction

Observe the alphabet represented by white dots and line segments in the figure. Starting from 'A', what is the second letter composed of eight white dots? Output this letter in uppercase.

Ground Truth

G

Figure 12. **A perception-level problem example from the MM-IFEval benchmark in the poster category.**



Instruction

Sam and Tom used the red box and Tom used the blue box. They each gave three answers. Would you please judge which of the two boys found more differences? Print the name of the the winning boy directly.

Ground Truth

Tom

Figure 13. **A perception-level problem example from the MM-IFEval benchmark in the finding difference category.**

Instruction generation prompt

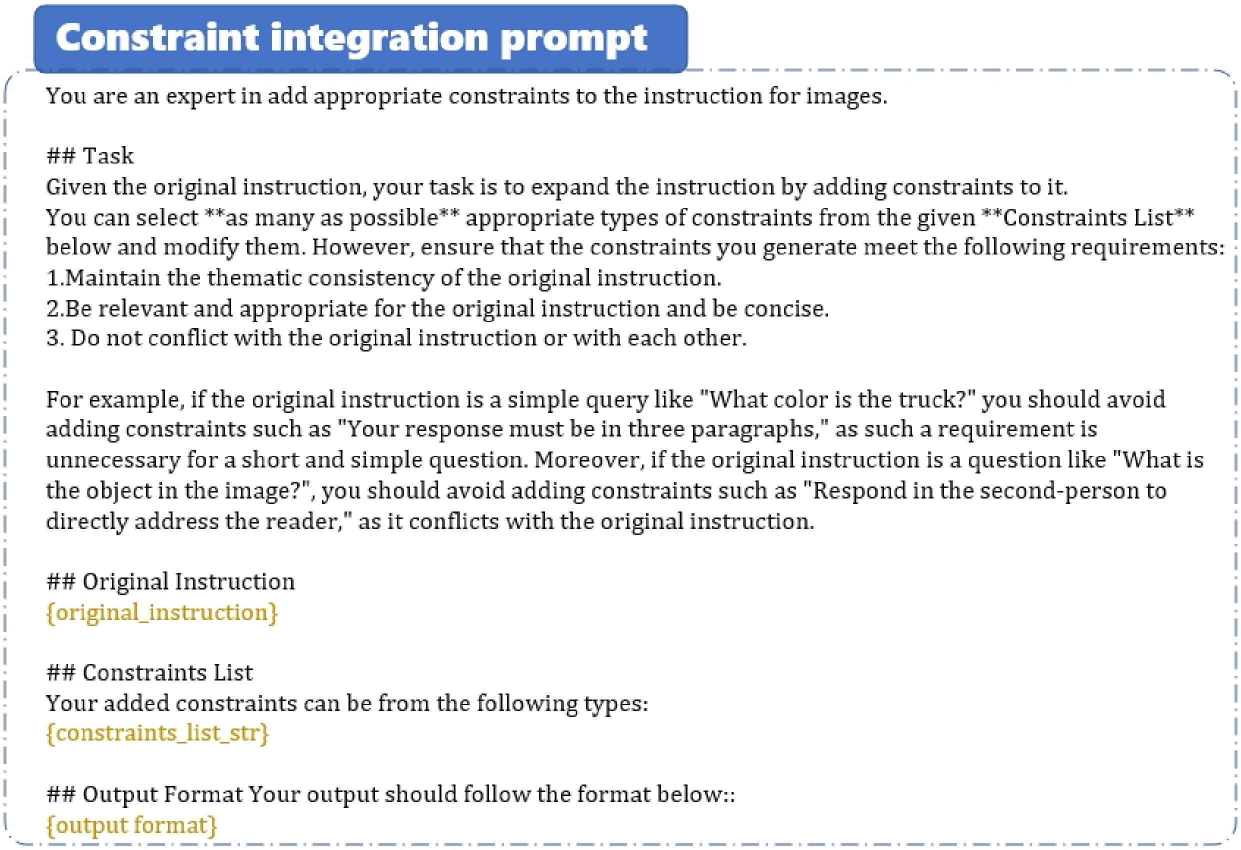
You are an expert in generating concise instructions for images.

Task
Given the image, generate a list of appropriate instructions for it.
Your instructions should not be too long or overly detailed, and they should not include any specific details about the image.
On one hand, you can choose appropriate instructions cases for the provided image from the Examples and modify them naturally for the image.
On the other hand, you can generate new instructions, but only if these new instructions are relevant and appropriate for the image.

Examples
{original instructions list}

You output format should be in the following format:
{output format}

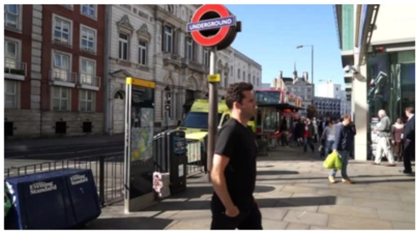
Figure 14. Prompt template for image generation instructions using a large language model in MM-IFEngine.



Constraint integration prompt

```
You are an expert in add appropriate constraints to the instruction for images.

## Task
Given the original instruction, your task is to expand the instruction by adding constraints to it.
You can select **as many as possible** appropriate types of constraints from the given **Constraints List**
below and modify them. However, ensure that the constraints you generate meet the following requirements:
1.Maintain the thematic consistency of the original instruction.
2.Be relevant and appropriate for the original instruction and be concise.
3. Do not conflict with the original instruction or with each other.

For example, if the original instruction is a simple query like "What color is the truck?" you should avoid
adding constraints such as "Your response must be in three paragraphs," as such a requirement is
unnecessary for a short and simple question. Moreover, if the original instruction is a question like "What is
the object in the image?", you should avoid adding constraints such as "Respond in the second-person to
directly address the reader," as it conflicts with the original instruction.

## Original Instruction
{original_instruction}

## Constraints List
Your added constraints can be from the following types:
{constraints_list_str}

## Output Format Your output should follow the format below::
{output format}
```

Figure 15. prompt template for integrating constraints in MM-IFEngine.



Instruction

Write a short poem that captures the hustle and bustle of city life depicted in this image.

Constraints

1.Write each stanza so it includes a parenthetical phrase.
2.Each stanza should have 4 lines.
3.Use **bold** for every noun.
4.Start the poem with 'In the heart of the city...' and end with '...the city never sleeps.'
5.Use present tense throughout the poem.
6.Use the third-person perspective to maintain an objective viewpoint.

Answer

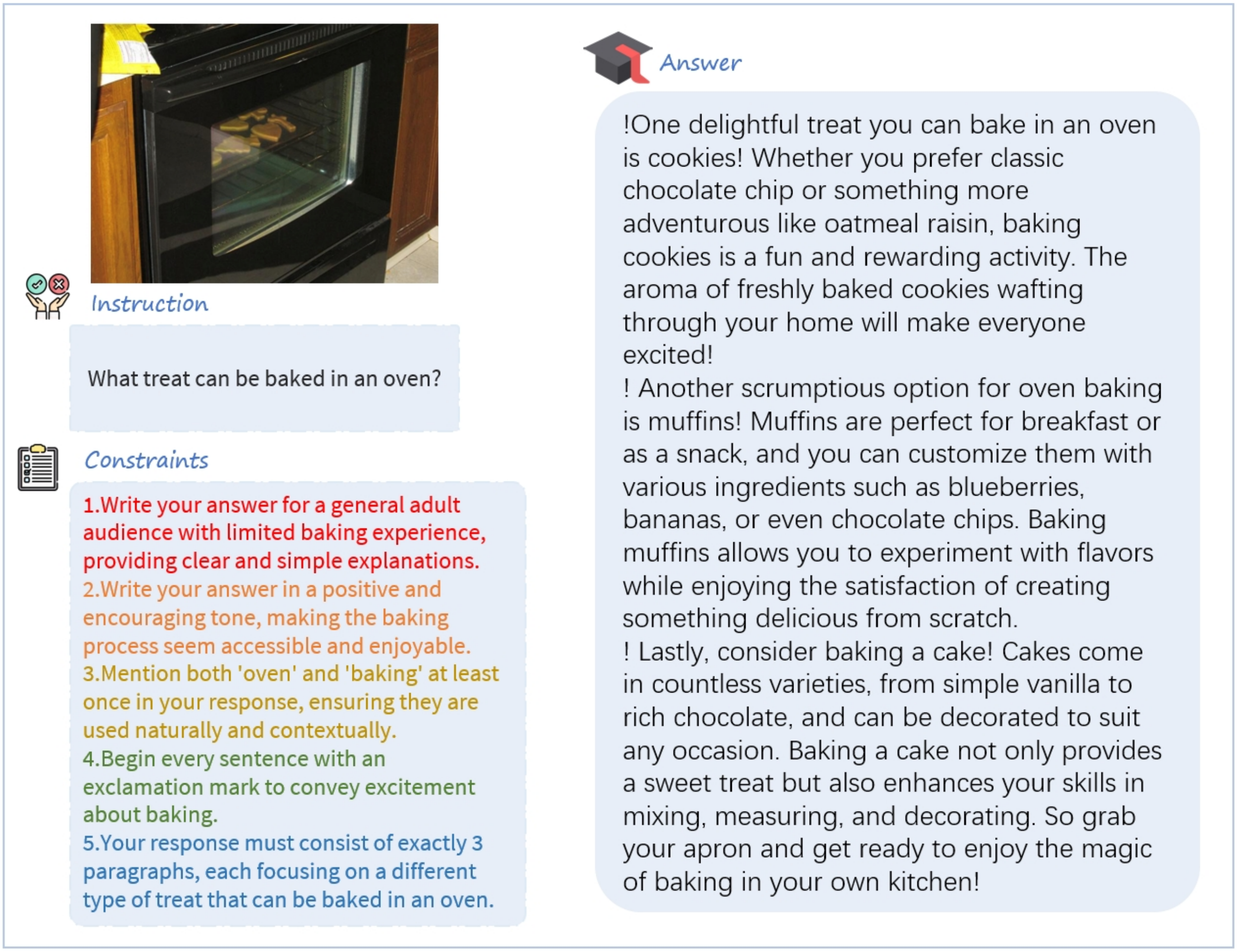
In the heart of the city, where the **streets** hum with life,
People rush (beneath the bright blue sky),
Buses pass (**tires** whispering on the road),
Every moment feels like a lively show.

The **signs** stand tall amidst the crowd,
An ambulance speeds by (with urgent sound),
Newspapers rest (in their quiet stack),
As footsteps weave through the urban track.

Buildings rise, their facades grand,
A man walks past (across the land),
Voices blend into ambient noise,
Life unfolds without any poise.

Shoppers carry bags (full of delight),
Sunlight dances (on the pavement's sight),
Each face tells tales of daily grind,
Yet spirits soar, unconfined....the city never sleeps.

Figure 16. A sample constructed by MM-IFEngine pipeline from cc3m dataset



Instruction

What treat can be baked in an oven?

Constraints

1.Write your answer for a general adult audience with limited baking experience, providing clear and simple explanations.
2.Write your answer in a positive and encouraging tone, making the baking process seem accessible and enjoyable.
3.Mention both 'oven' and 'baking' at least once in your response, ensuring they are used naturally and contextually.
4.Begin every sentence with an exclamation mark to convey excitement about baking.
5.Your response must consist of exactly 3 paragraphs, each focusing on a different type of treat that can be baked in an oven.

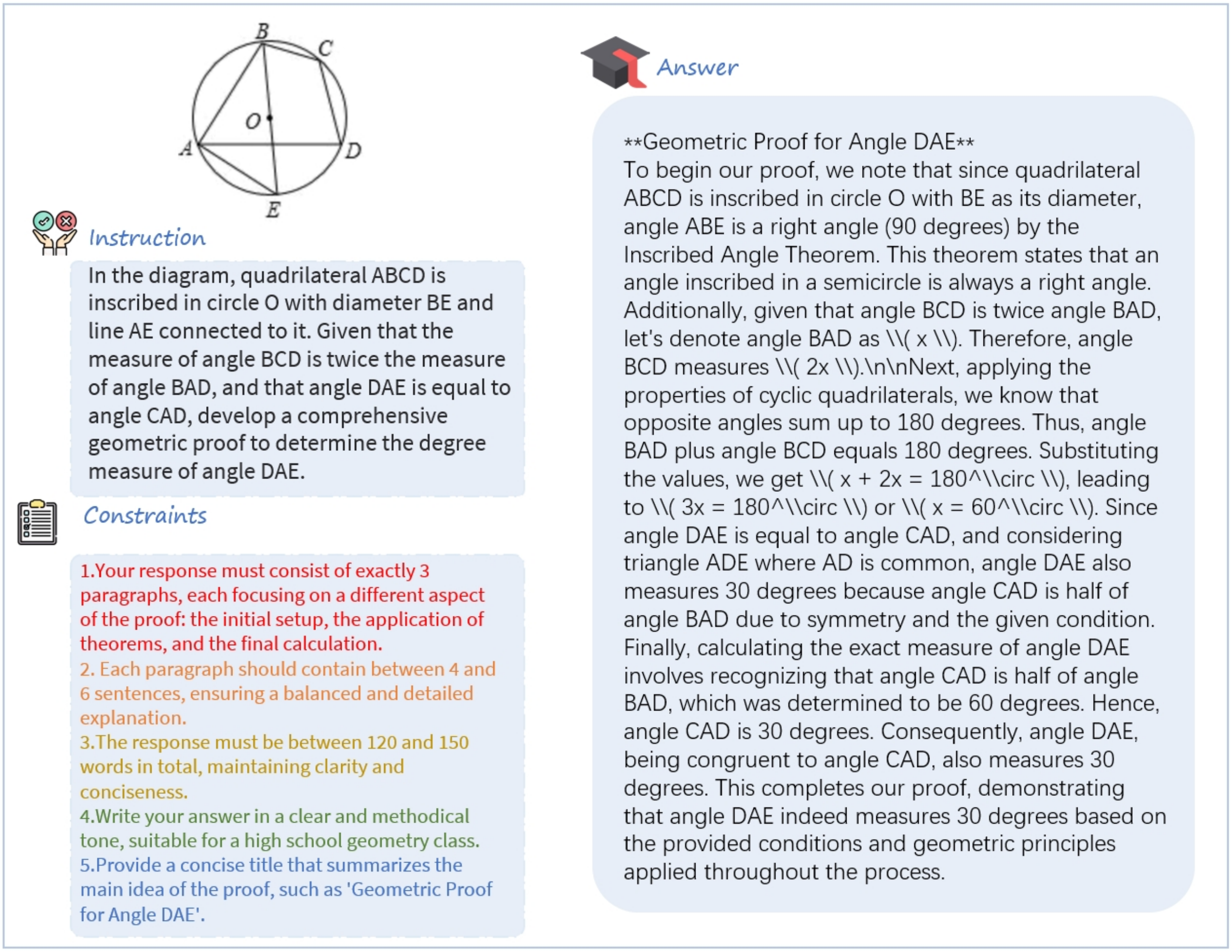
Answer

!One delightful treat you can bake in an oven is cookies! Whether you prefer classic chocolate chip or something more adventurous like oatmeal raisin, baking cookies is a fun and rewarding activity. The aroma of freshly baked cookies wafting through your home will make everyone excited!

! Another scrumptious option for oven baking is muffins! Muffins are perfect for breakfast or as a snack, and you can customize them with various ingredients such as blueberries, bananas, or even chocolate chips. Baking muffins allows you to experiment with flavors while enjoying the satisfaction of creating something delicious from scratch.

! Lastly, consider baking a cake! Cakes come in countless varieties, from simple vanilla to rich chocolate, and can be decorated to suit any occasion. Baking a cake not only provides a sweet treat but also enhances your skills in mixing, measuring, and decorating. So grab your apron and get ready to enjoy the magic of baking in your own kitchen!

Figure 17. A sample constructed by MM-IFEngine pipeline from Allava dataset



Instruction

In the diagram, quadrilateral ABCD is inscribed in circle O with diameter BE and line AE connected to it. Given that the measure of angle BCD is twice the measure of angle BAD, and that angle DAE is equal to angle CAD, develop a comprehensive geometric proof to determine the degree measure of angle DAE.

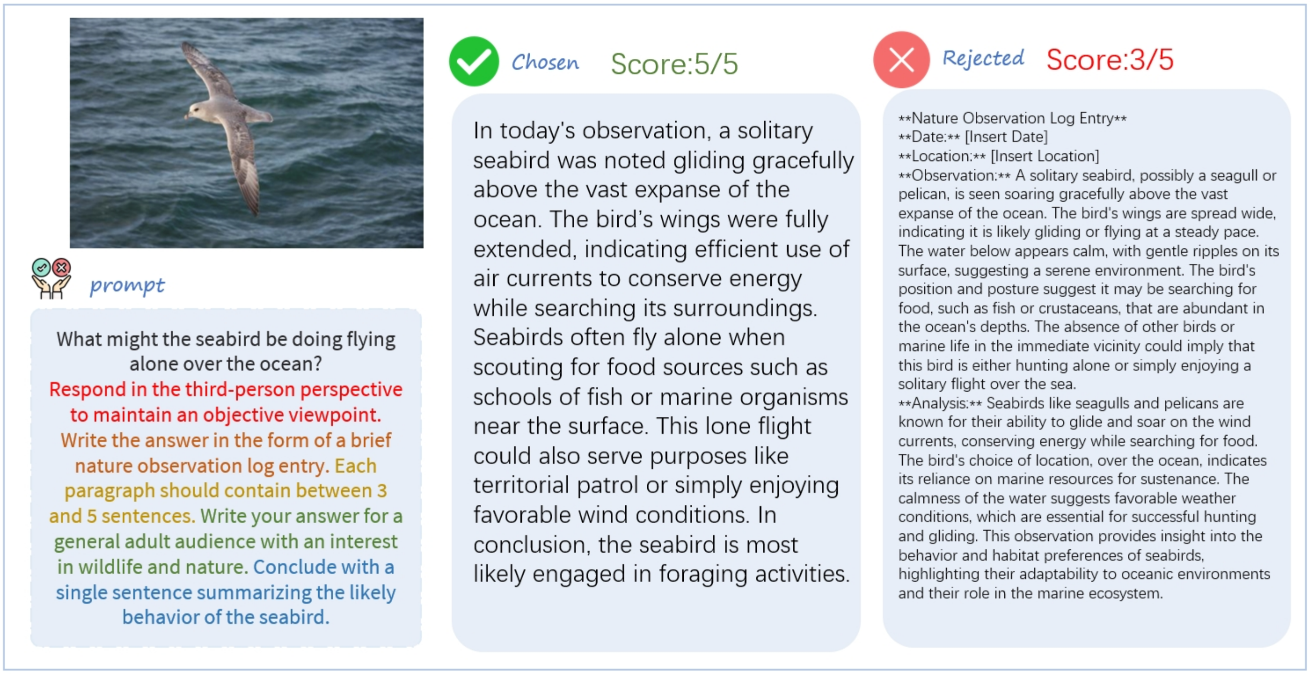
Constraints

1.Your response must consist of exactly 3 paragraphs, each focusing on a different aspect of the proof: the initial setup, the application of theorems, and the final calculation.
2. Each paragraph should contain between 4 and 6 sentences, ensuring a balanced and detailed explanation.
3.The response must be between 120 and 150 words in total, maintaining clarity and conciseness.
4.Write your answer in a clear and methodical tone, suitable for a high school geometry class.
5.Provide a concise title that summarizes the main idea of the proof, such as 'Geometric Proof for Angle DAE'.

Answer

Geometric Proof for Angle DAE
To begin our proof, we note that since quadrilateral ABCD is inscribed in circle O with BE as its diameter, angle ABE is a right angle (90 degrees) by the Inscribed Angle Theorem. This theorem states that an angle inscribed in a semicircle is always a right angle. Additionally, given that angle BCD is twice angle BAD, let's denote angle BAD as \\(x \\). Therefore, angle BCD measures \\(2x \\).\n\nNext, applying the properties of cyclic quadrilaterals, we know that opposite angles sum up to 180 degrees. Thus, angle BAD plus angle BCD equals 180 degrees. Substituting the values, we get \\(x + 2x = 180^\\circ \\), leading to \\(3x = 180^\\circ \\) or \\(x = 60^\\circ \\). Since angle DAE is equal to angle CAD, and considering triangle ADE where AD is common, angle DAE also measures 30 degrees because angle CAD is half of angle BAD due to symmetry and the given condition. Finally, calculating the exact measure of angle DAE involves recognizing that angle CAD is half of angle BAD, which was determined to be 60 degrees. Hence, angle CAD is 30 degrees. Consequently, angle DAE, being congruent to angle CAD, also measures 30 degrees. This completes our proof, demonstrating that angle DAE indeed measures 30 degrees based on the provided conditions and geometric principles applied throughout the process.

Figure 18. A sample constructed by MM-IFEngine pipeline from geo170k dataset


prompt
Chosen
Score:5/5
Rejected
Score:3/5

What might the seabird be doing flying alone over the ocean? Respond in the third-person perspective to maintain an objective viewpoint. Write the answer in the form of a brief nature observation log entry. Each paragraph should contain between 3 and 5 sentences. Write your answer for a general adult audience with an interest in wildlife and nature. Conclude with a single sentence summarizing the likely behavior of the seabird.

In today's observation, a solitary seabird was noted gliding gracefully above the vast expanse of the ocean. The bird's wings were fully extended, indicating efficient use of air currents to conserve energy while searching its surroundings. Seabirds often fly alone when scouting for food sources such as schools of fish or marine organisms near the surface. This lone flight could also serve purposes like territorial patrol or simply enjoying favorable wind conditions. In conclusion, the seabird is most likely engaged in foraging activities.

Nature Observation Log Entry
Date: [Insert Date]
Location: [Insert Location]
Observation: A solitary seabird, possibly a seagull or pelican, is seen soaring gracefully above the vast expanse of the ocean. The bird's wings are spread wide, indicating it is likely gliding or flying at a steady pace. The water below appears calm, with gentle ripples on its surface, suggesting a serene environment. The bird's position and posture suggest it may be searching for food, such as fish or crustaceans, that are abundant in the ocean's depths. The absence of other birds or marine life in the immediate vicinity could imply that this bird is either hunting alone or simply enjoying a solitary flight over the sea.
Analysis: Seabirds like seagulls and pelicans are known for their ability to glide and soar on the wind currents, conserving energy while searching for food. The bird's choice of location, over the ocean, indicates its reliance on marine resources for sustenance. The calmness of the water suggests favorable weather conditions, which are essential for successful hunting and gliding. This observation provides insight into the behavior and habitat preferences of seabirds, highlighting their adaptability to oceanic environments and their role in the marine ecosystem.

Figure 19. A DPO training set sample, where the rejected data is obtained by removing 33% of the constraints

Prompt to choose verified function and extract params

Please analyze the following constraint and select the most appropriate function from the given list to verify this constraint. Then extract the required parameters for the verification function from the constraint.

Constraint content: {constraint value}

Available verification functions:
{all candidate validation function names and parameters}

Please complete the analysis following these steps:
Your task:
1. Select the most appropriate verification function from the above list (return empty if none is suitable)
2. Extract the required parameters from the constraint based on the function description

Please return the result in JSON format as follows:
{output format}

Figure 20. Prompt template for automated verification function selection and paramater extraction

Compare Judge Prompt

You are an expert in judging whether the respone follow the given constraint. Your task is to assess whether the model's response satisfies the given constraint and return True or False. I will provide you with the constraint and the model's response under this constraint. To assist with your evaluation, I will also provide you with the model's response to the same question without the constraint.
Constraint: {constraint}
Response under the constraint: {pred_with_constraint}
Response without the constraint: {pred_without_constraint}

Please follow the steps below to evaluate:
Step 1. Compare the model's response under the constraint with its response without the constraint. If you believe these two answers are very similar, it means the model has not fully considered the impact of the constraint on the answer. Please return False.
Step 2. Compare the model's response under the constraint with the content of the constraint. If you believe the model's response does not meet the requirements specified in the constraint, return False. Otherwise, if the response effectively satisfies the constraint, return True.
Response Format: Your answer should only include "True" or "False", and no additional text.
.

Figure 21. Prompt template for Compare Judge Method

Direct Judge Prompt

Your task is to evaluate whether the response from an AI assistant adheres to all of the given constraints.
Please follow the requirements below to make the judgment:
1. Be strict and consistent in your assessment.
2. You should refer to the content of image to make the judgment.
3. For one constraint, if the response fails to fully meet the constraint, give it a score of 0. Otherwise, give it a score of 1. <start of response>
{prediction}
<end of response>
<start of constraint list>
{constraints_str}
<end of constraint list>
You should judge and explain for each constraint in the constraint list without omitting any constraint. Finally, list scores of all the constraints in one sentence.
You should strictly follow the format below:
Judgement: ...
Summary: Score of constraint_1: x/1, Score of constraint_2: x/1, Score of constraint_3: x/1, ..., Score of constraint_n: x/1.

Figure 22. Prompt template for Direct Judge Method

Main Class	Subclass	Evaluation	Description	Example
A. Rhetoric & Logic	A.1 Rhetoric requirements	Compare Judge	Constraint that requires the response to use a specific rhetorical technique.	"Your output should include a metaphor."
	A.2 Logical relation	Direct Judge	Constraint that ensures logical cohesion within the response by requiring specific logical connectors or structures.	"Each paragraph must contain at least one cause-and-effect relationship."
B. Format limit	B.1 Natural language	Direct Judge	Constraint specifying which natural language(s) should be used in the response.	"Please answer in Spanish."
	B.2 Part of speech	Direct Judge	Constraint that requires the response to use a specific part of speech.	"Use at least three adjectives in your response."
	B.3 Sentence structure	Direct Judge	Constraint that specifies special sentence structures to be used in the response.	"Write each sentence so it includes a parenthetical phrase."
	B.4 Tense requirements	Direct Judge	Constraint that specifies the use of multiple tenses within the response.	"In past tense totally."
	B.5 Punctuation	Rule-base	Constraint specifying unconventional yet feasible punctuation usage in the response.	"Replace all periods with semicolons."
	B.6 Highlight	Direct Judge	Constraint that specifies a unique but manageable method for highlighting text.	"Use **bold** for every noun."
	B.7 Title requirements	Direct Judge	Constraint that specifies how titles should be added to the response.	"Provide a concise title that summarizes the main idea."
	B.8 Style requirements	Compare Judge	Constraint that specifies an unconventional or distinctive writing style for the response.	"Write the answer in the form of a brief detective story."
	B.9 Case requirements	Direct Judge	Constraint specifying an unusual yet readable approach to letter case in the response.	"Write all nouns in UPPERCASE and all adjectives in lowercase."
	B.10 Unstrict format	Direct Judge	Constraint specifying a unique format for the output while keeping it approachable.	"Format your response as a short play script with speaker labels."
	B.11 Strict format	Direct Judge	Constraint that requires the response to follow a strictly defined format.	"Please provide the output as well-formed XML with custom tags."
	B.12 Number and List	Direct Judge	Constraint for using numbered or bulleted lists in the response.	"Present all key points as a numbered list with bulleted sub-lists."
	B.13 Wrap up	Direct Judge	Constraint that requires a concise, well-structured summary or conclusion.	"Provide a final paragraph summarizing the key arguments."
	B.14 First letter	Direct Judge	Constraint specifying a pattern for the first letters of sentences or paragraphs.	"Each sentence should begin with a letter that progresses through the alphabet."
C. Text Length limit	C.1 Paragraph limit	Rule-base	Constraint that specifies the number of paragraphs in the response.	"Your response must consist of exactly 4 paragraphs."
	C.2 Sentence limit	Rule-base	Constraint that specifies the number of sentences in each paragraph.	"Totally use 5 sentences in your response."
	C.3 Word limit	Rule-base	Constraint that specifies a small range for the total number of words in the text.	"Your response must be a single word or phrase."
D. Math limit	D.1 Precision	Rule-base	Constraint that specifies the level of precision required in mathematical calculations.	"Keep two decimal places for all numbers in the answer."
	D.2 Scientific notation	Rule-base	Constraint that requires the use of scientific notation for large or small numbers.	"Express all numbers greater than 1,000 in scientific notation."
E. Action limit	E.1 Role imitation	Compare Judge	Constraint requiring the response to imitate the tone and style of a specific role or public figure.	"Please answer in the style of a sports commentator."
	E.2 Prefix and Suffix	Rule-base	Constraint that requires the response to begin or end with a specific phrase or symbol.	"Please start your answer with 'Once upon a time...'."
	E.3 Tone requirement	Compare Judge	Constraint specifying an emotional tone for the response.	"Write your answer in a positive and encouraging tone."
	E.4 Perspective	Direct Judge	Constraint that specifies a narrative perspective for the response.	"Write your answer in the first-person singular as a personal account."
	E.5 Target audience	Compare Judge	Constraint requiring the response to be tailored for a specific audience.	"Craft your response as if explaining to high school students."
	E.6 Situation	Compare Judge	Constraint requiring the response to be set in a specific situation or scenario.	"Answer as if you are giving safety instructions before a flight."
	E.7 Prior condition	Direct Judge	Constraint stating that when a specific condition is met, the response must follow a particular process.	"If the user requests legal advice, begin with a disclaimer."
F. Keyword	F.1 Mention	Rule-base & Direct Judge	Constraint that requires including a specific keyword a certain number of times.	"Mention 'GreenTech' exactly three times throughout."
	F.2 Not mention	Rule-base & Direct Judge	Constraint that requires avoiding specific keywords or phrases.	"Do not mention the words 'budget' or 'investment'."
	F.3 Multiple mention	Rule-base & Direct Judge	Constraint requiring including multiple specified keywords in a balanced manner.	"Mention both 'sustainability' and 'renewable energy' at least twice."
	F.4 Keyword variation	Direct Judge	Constraint requiring the use of synonyms or variations of a given keyword.	"Use at least three synonyms for 'innovation' throughout your text."

Table 5. Constraint Categories and Evaluation Methods for MM-IFEval

Category	Instruction
Descriptive Analysis	Describe the animal's typical habitat, diet, and one unique behavioral trait.
	Provide a detailed analysis of the image, including the setting, characters, and notable objects.
	Explain the activity taking place in the image.
	Describe the activities of the person on the left in the image.
Emotional & Perspective	What emotions do you think the person in this image might be feeling?
	Imagine you are the person on the left in the scene depicted in this image, write a story about what you would do next.
	Personify the sign in the image and express its feelings about the rule it presents.
Creative Writing	Create a short conversation between any two individuals in the scene.
	Pretend this snapshot belongs to a larger story. Write a quick paragraph setting up the next plot twist.
	Use this picture as your muse. Craft a brief poem—any style—that captures the emotion you sense.
	Turn this scene into a short children's story focusing on wonder and curiosity.
	Write a short poem with two stanzas, inspired by the emotion or content depicted in this image.
Social Media & Content	Assume this is an image you are about to post on Twitter. Please provide a short, upbeat caption describing it.
	Assume you are creating a Pinterest pin with this image. Write a short inspirational or motivational caption to accompany it.
	If this image were promoting an upcoming event, compose a quick announcement with the date, a highlight of what to expect, and a call-to-action.
Role Play	Imagine you are the photographer who took this picture. Briefly explain why you chose to capture this particular moment and what story you hope it conveys.

Table 6. Task Pool for MM-IFEngine

Verified Function Name	**Function Parameters**	**Constraint Example**	**Parameter Example**
check_whether_response_paragraph_number_in_range	lower_bound:int, upper_bound:int	The number of text paragraphs be at least 3	[3, 10000]
check_whether_response_sentence_number_in_range	lower_bound:int, upper_bound:int	The number of sentences be exactly 3	[3, 3]
check_whether_each_paragraph_sentence_number_in_range	lower_bound:int, upper_bound:int	The number of sentences in each paragraph be less than 3	[0, 2]
check_whether_each_paragraph_sentence_number_in_range_list	ranges:List[tuple]	The number of sentences in the first paragraph be exactly 3, and in the second paragraph be at most 2	[[(3, 3), (1, 2)]]
check_whether_each_paragraph_sentence_number_exceeds	exceed_num:int, upper_bound:int	Each new paragraph should have 1 sentence more than the previous one, no paragraph exceeds 7 sentences	[1, 7]
check_whether_response_word_count_in_range	lower_bound:int, upper_bound:int	The number of words should be between 50 and 80	[50, 80]
check_whether_each_paragraph_word_count_in_range	lower_bound:int, upper_bound:int	The number of words in each paragraph should be between 50 and 80	[50, 80]
check_whether_each_paragraph_word_count_in_range_list	ranges:List[tuple]	The number of words in the first paragraph be between 20 and 30, in the second between 50 and 80	[[(20, 30), (50, 80)]]
check_whether_whole_response_not_contain_certain_substring	substring:str	The response should not contain the word ”apple”	[”apple”]
check_whether_whole_response_not_contain_certain_substrings	substrings:List[str]	The response should not contain the words ”apple” and ”banana”	[[”apple”, ”banana”]]
check_whether_each_sentence_begin_with_certain_substring	substring:str	Each sentence should start with exclamation point	[”!”]
check_whether_each_sentence_end_with_certain_substring	substring:str	Each sentence should end with ”apple”	[”apple”]
check_whether_whole_response_begin_with_certain_substring	substring:str	The response should start with ”apple”	[”apple”]
check_whether_whole_response_end_with_certain_substring	substring:str	The response should end with ”apple”	[”apple”]
check_whether_keywords_metioned_in_range	keywords:List[str], lower_bound_times:int, upper_bound_times:int	The response should mention the word ”apple” at least 3 times	[[”apple”], 3, 10000]
check_number_precision_in_response	precision:int	The numbers in the response should have 2 decimal places	[2]
check_whether_has_no_number_in_response	-	The response should not contain any number	[]
check_scientific_notation_precision_in_response	significant_digits:int	The numbers in the response should have 3 significant digits	[3]

Table 7. Verification Functions for rule-based evaluation method in MM-IFEval