

Appendix For “TimeCausality: Evaluating the Causal Ability in Time Dimension for Vision Language Models”

A. Prompts used in our TimeCausality

A.1. Prompt and Models for Dataset Generation

We use the Grounded-Segment-Anything¹ as the Auto-Label Model, and the GPT-4o as the LLM and the inpainting model, the prompt used for GPT-4o is shown in Tab S1.

A.2. Prompt for LLM-as-a-Judge

For the Aspect II and the Aspect III, we use the LLM to judge the models’ output, the prompt used for evaluation the reasoning quality is shown in Tab S2.

A.3. Prompt for Evaluation

For each Aspect, we use different prompts for VLMs to evaluate. The prompt for Aspect I evaluation is shown in Tab S3, and the prompt for Aspect II and Aspect III are shown in Tab S4.

B. Human Verification.

To ensure the quality of our TimeCausality, every sample in our benchmark is verified by human, the guidelines for check are shown in Tab S5.

C. More Cases in TimeCausality

In this section, as shown in Fig S1, we present additional examples from various categories in TimeCausality. Each sample consists of an image pair, an object name, an edit instruction, a type, reasoning rationales, and inferring rationales. Notably, reversing the order of the image pair can significantly affect the model’s inference. Many models exhibit strong positional bias, often assuming that the image appearing earlier (e.g., the upper image) corresponds to the earlier event. To mitigate the impact of such biases during evaluation, we include the F1 score as an important reference metric.

D. Experiments Details.

For the Closed-Source VLMs, we revoke the official api for evaluation. For the Open-Source VLMs, we use their corresponding weight on Huggingface², with the package³ to deploy them locally with the default setting. All locally deployed experiments are conducted on a server with NVIDIA RTX4090s.

E. The Gap Between Aspect I with II&III

In our TimeCausality, we define three distinct aspects, each corresponding to a different evaluation dimension. Correct performance in one aspect does not necessarily imply correct performance in the others. As illustrated in Fig. S2, GPT-4o correctly identifies the temporal order between the two images in the first task (top of the figure). However, when the order of the images is reversed, GPT-4o produces an incorrect prediction (middle of the figure). Finally, when prompted to provide an explanation, GPT-4o gives a reasonable rationale—inferring the temporal order from the degree of fruit decay—yet still outputs the wrong temporal judgment. This case highlights the inconsistency between reasoning and final decision in GPT-4o, which further motivates our design of evaluating models from three separate aspects.

¹<https://github.com/IDEA-Research/Grounded-Segment-Anything>

²<https://huggingface.co/>

³<https://github.com/InternLM/lmdeploy>

Example prompt for data generation

Based on the provided object describe an irreversible transformation it could undergo. The transformation should fit one of the following categories:

1. Physical change: Structural or morphological changes caused by external forces or environmental factors (e.g., a glass cup shattered).
2. Chemical change: Changes to the object's chemical composition, resulting in new substances (e.g., food rotting).
3. Natural phenomenon: Changes due to natural processes or the passage of time (e.g., aging of a person).
4. Environmental modification: Alterations caused by human intervention over time (e.g., a historical building renovated into a modern structure).
5. Artificial processing: Transformations caused by human-made techniques or craftsmanship (e.g., raw clay turned into pottery).

For each input object:

- Provide a clear edit instruction for the irreversible transformation.
- Justify why this transformation logically follows from the original state and cannot revert.
- Explain how this change could occur, referencing plausible causes or processes.
- Specify the type of the transformation.

Output your response in the following structured format:

<type>

[Specify the type of change, e.g., "Physical change", "Chemical change", "Natural phenomenon", "Environmental modification", or "Artificial processing."]

</type>

<edit_instruction>

[Provide a concise and clear directive for the transformation, e.g., "Transform the glass cup into shattered pieces scattered across a surface."]

</edit_instruction>

<rationales>

[Explain why the modified image depicts a later state in time or process compared to the original, e.g., "A shattered glass cannot return to its original intact state, making this an irreversible transformation."]

</rationales>

<operation>

[Describe what could cause the change to occur, e.g., "The glass was likely subjected to a sudden external force or impact."]

</operation>

Examples:

Input: A glass cup

<type>Physical change</type>

<edit_instruction> Transform the glass cup into shattered pieces scattered across a surface. </edit_instruction>

<rationales>A shattered glass cannot return to its original intact state, making this an irreversible transformation.</rationales>

<operation>The glass was likely subjected to a sudden external force or impact.</operation>

Input: A slice of bread

<type>Chemical change</type>

<edit_instruction> Show the bread as moldy and green with visible spores on its surface. </edit_instruction>

<rationales>Once bread molds, the chemical composition changes, and it cannot revert to its original edible state.</rationales>

<operation>The bread likely underwent chemical reactions caused by exposure to moisture and microbial activity over time.</operation>

Input: A person

<type>Natural phenomenon</type>

<edit_instruction> Depict the person as elderly, with wrinkles, gray hair, and visible signs of aging. </edit_instruction>

<rationales>Aging is a natural, irreversible process that occurs over time and cannot be reversed to a previous state.</rationales> <operation>The changes occur naturally due to biological aging processes over years.</operation>

Input: A small village

<type>Environmental modification</type>

<edit_instruction> Transform the small village into a modern urban area with tall buildings and paved roads. </edit_instruction>

<rationales>Urbanization is an irreversible process that permanently alters the original layout and structure of the village.</rationales> <operation>This change could occur due to long-term urban development and modernization initiatives by humans.</operation>

Now, for the following input object, apply the same structure:

Input: {Input Source}

Table S1. Example prompt for data generation

Example prompt for LLM-as-a-Judge

You will be given a ground truth and model output couple.

Your task is to provide a 'total rating' scoring how well the model output matches the semantic meaning of the ground truth.

Give your answer as an integer on a scale of 0 to 5, where 0 means that the model output is completely unrelated to the ground truth, and 5 means that the model output perfectly matches the semantic meaning of the ground truth.

Provide your feedback as follows:

Feedback:

Total rating: (your rating, as an integer between 0 and 5)

Now here are the ground truth and model output.

Ground Truth: {ground_truth}

Model Output: {model_output}

Feedback:

Total rating:

Table S2. Example prompt for LLM-as-a-Judge

Example prompt for Aspect I Evaluation

Question: In the given image, the bottom and top sides depict the states of the same {Source Object} at different points in time. Based on the progression of time, which sequence is correct?

Only consider natural, long-term changes (e.g., aging, oxidation, rust, or physical degradation). Ignore any possibilities of digital alterations, retouching, or external modifications.

Choices:

A. From bottom to top (Bottom happens first, top happens later).

B. From top to bottom (Top happens first, bottom happens later).

Output only in a single letter. (A or B) .

Table S3. Example prompt for Aspect I Evaluation

Example prompt for Aspect II & III Evaluation

Why does the {Source Object} in the scene on the top occur after the one on the bottom?

Answer in one sentence.

What caused the transition from the {Source Object} on the bottom to the {Source Object} on the top? Answer in one sentence.

Table S4. Example prompt for Aspect II & III Evaluation

Guidelines for Human Verification:

This is an image pair showing two states of the same scene or object. Please evaluate the following:

Temporal Order: Do these two images have a clear sequential relationship? In other words, does the content of one image obviously appear before or after the other?

Content Consistency: Do the two images maintain overall consistency? Can you confirm that they depict the same scene or the same object?

Reasoning Rationales: Does the reason for the order of these two images align with the provided Reasoning Rationales?

Inferring Rationales: Does the reason for the order of these two images align with the provided Inferring Rationales?

Final Check: Please also assess whether this image pair contains any disturbing or NSFW (Not Safe For Work) content, such as horror, gore, or other inappropriate material.

Table S5. Guidelines for Human Verification

Image pair Object Name Edit Instruction Type Reasoning Rationales Inferring Rationales

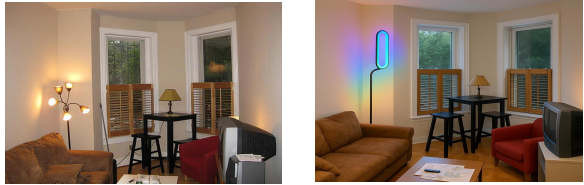



Image pair	Object Name	Edit Instruction	Type	Reasoning Rationales	Inferring Rationales
	wall lamp	Artificial processing			
		<p>Transform the wall lamp into a modern LED-based smart lamp with integrated technology.</p>			
		<p>Reasoning (Why): Once the wall lamp has been transformed into a smart lamp with integrated LED technology and smart features, it cannot revert to its original form as a traditional wall lamp without a complete redesign and removal of the modern components.</p>			
		<p>Inferring (What): This change could occur through human intervention by retrofitting the existing wall lamp with LED technology and incorporating smart functionalities, such as Wi-Fi connectivity and voice control, through electronics and software integration.</p>			
	man tennis player	Natural phenomenon			
		<p>Depict the man tennis player as elderly, showing signs of aging such as wrinkles, gray hair, and a slower physical demeanor.</p>			
		<p>Reasoning (Why): Aging is a natural, irreversible process that affects everyone over time, and the physical changes associated with aging cannot be undone.</p>			
		<p>Inferring (What): This transformation occurs naturally as a result of biological aging processes that take place over several decades.</p>			
	plane	Physical change			
		<p>Transform the plane into twisted and broken wreckage scattered across a field.</p>			
		<p>Reasoning (Why): Once a plane is destroyed in a crash, it cannot be restored to its original flying condition, making this an irreversible transformation.</p>			
		<p>Inferring (What): The plane likely experienced a catastrophic failure, leading to a crash caused by mechanical malfunction, human error, or severe weather conditions.</p>			
	motorbike	Environmental modification			
		<p>Transform the motorbike motorcycle into a rusted and abandoned state, with parts missing and overgrown with vegetation.</p>			
		<p>Reasoning (Why): Once a motorbike is left unused and exposed to the elements over time, it will undergo degradation that cannot be reversed to restore it to its original functional condition without significant repair or replacement of parts.</p>			
		<p>Inferring (What): This change could occur due to long-term neglect and exposure to environmental factors such as rain, humidity, and plant growth, leading to rusting and deterioration.</p>			


Figure S1. More Cases in Our Time Causality

User

Question: In the given image, the bottom and top sides depict the states of the same pear at different points in time. Based on the progression of time, which sequence is correct?

Only consider natural, long-term changes (e.g., aging, oxidation, rust, or physical degradation). Ignore any possibilities of digital alterations, retouching, or external modifications.

Choices: A. From bottom to top B. From top to bottom Output only in a single letter. (A or B)



GPT4o

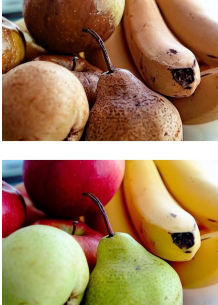
B

User

Question: In the given image, the bottom and top sides depict the states of the same pear at different points in time. Based on the progression of time, which sequence is correct?

Only consider natural, long-term changes (e.g., aging, oxidation, rust, or physical degradation). Ignore any possibilities of digital alterations, retouching, or external modifications.

Choices: A. From bottom to top B. From top to bottom Output only in a single letter. (A or B)



GPT4o

B

User

Explain the reason to me

GPT4o

The correct answer is B (From top to bottom) because:
 The top half of the image shows fruits that appear aged or overripe. The bananas have dark spots and browning ends, and the pears are mottled with brown, showing signs of oxidation and decay.
 The bottom half of the image shows the same types of fruit (bananas, pears, apples), but they look fresh and vibrant. The bananas are a more even yellow with fewer dark spots, the pears are firm and green, and the apples are bright red.
 This visual progression—from fresh (bottom) to aged/decayed (top)—is a natural result of time and fruit ripening followed by spoilage, indicating the sequence is B: from top to bottom.

Figure S2. What makes VLMs fail to solve our proposed task? In this case, GPT4o successfully identifies the temporal order of in the first image-pair, but fails to do so on the reversed image. This shows an important point in the Aspect I evaluation: consistency. At the same time, for the reversed image-pair, GPT4o gives a completely correct reasoning process, but gives a wrong answer. This shows that although VLM may have certain capabilities in Aspect II&III, it does not successfully associate them with visual information, resulting in the wrong final answer.