CaST-Bench: Benchmarking Causal Chain-Grounded Spatio-Temporal Reasoning for Video Question Answering

Supplementary Material

8. Appendix Overview

The organization of the appendix is as follows:

Sec. 9: CaST-Bench Samples from Each Category
Sec. 10: Details of Data Annotation Pipeline
Sec. 11: Details of Experiment Setup
Sec. 13: Case Studies and Failure Analysis
Sec. 14: Social Impact, License, and Access

9. CaST-Bench Samples from Each Category

Due to page limitation in the main manuscript, we show more examples regarding all question types, as follows.

- **Causal Explanation** Questions that explain the reasons (*why*) or mechanisms (*how*) behind actions or events.
  - Why questions (reasons), shown in Fig. 9
  - How questions (mechanisms), shown in Fig. 10
- **Counterfactual Reasoning** Questions that infer the consequence of a single alteration to a key causal element.
  - Physical counterfactual, shown in Fig. 11
  - Social counterfactual, shown in Fig. 12
- **Predictive Anticipation** Questions that require predicting the most probable and immediate outcome.
  - Behavioral anticipation, shown in
  - Physical counterfactual, shown in Fig. 13
- **Inferential Description** Questions that infer implicit attributes or states (e.g., roles, intentions, emotions).
  - Skill/expertise inference, shown in Fig. 14
  - ...

Detailed definition of each question type can be found in Sec. 10.4.

10. Details of Data Annotation Pipeline

10.1. Video Selection

Our benchmark targets causal reasoning in realistic, cluttered scenes where multiple actors interact over time. As discussed in Sec. 4, carefully curating the raw video pool is essential: studio footage or single-actor clips often lack the competing causal cues and spatial ambiguity needed to stress-test grounding. We therefore begin CaST-Bench construction with a *Video Selection* stage that selects complex activities from the SegmentAnything-Video (SAV) dataset [29], ensuring that downstream annotation stages receive videos containing diverse objects, overlapping motions, and stable visual quality.

We operationalize the above motivation with a tool that iterates over SAV clips matching a curator-specified pattern, submits each video to a frontier multimodal model via a standard API endpoint, and relies on the prompt (shown below) to approve a clip only when (a) multiple humans are visible and (b) the visible people are engaged in different actions, which directly encourages multi-instance causal chains; the same prompt rejects clips with motion artifacts (camera shake, rapid cuts or pans), low visibility, or severe compression because poor spatial alignment would confound later evidence annotation, and the resulting 1,015 accepted videos constitute the seed pool for the remaining stages of the pipeline.

Prompt 1: Automated Video Selection

You are a video analysis expert. Task - Watch the entire video. Return exactly one of: - True — the video shows multiple people engaged in different actions. - False — otherwise. Output format - Return only True or False (no quotes, no extra text, no punctuation). Decision rules 1) Human presence - If only one person is present, return False. - If multiple people are present and each visible person's behavior are differnt, return True. 2) Visual quality and stability - Return False if there is: a) strong camera shake; b) rapid viewpoint shifts or frequent cuts; c) very low image quality (subject not discernible, heavy blur, severe compression, lighting too dark or overexposed); d) fast camera movement (e.g., quick or whip pans); e) severe camera tilt. Remember: return only True or False.

10.2. Spatio-Temporal Fine-grained Instance Description

Generating detailed and accurate descriptions for specific object instances within complex, dynamic scenes presents a significant challenge. Vision-Language Models (VLMs) can be easily distracted by background clutter or other moving objects, leading to descriptions that are either inaccurate or lack focus on the intended instance. To address this, we developed a multi-stage pipeline that systematically isolates each instance and generates both static and dynamic descriptions, forming a rich corpus for the subsequent generation of causal questions. This process consists of three key stages: instance isolation for dynamic analysis, static description for contextual understanding, and dynamic description for capturing temporal behavior.

Stage 1: Instance Isolation for Dynamic Analysis The primary challenge in analyzing the behavior of a single object is isolating it from the surrounding environment. To enable a VLM to focus exclusively on the actions of one instance, we first create a cropped video clip that follows the target object throughout its appearance in the original video. Using the instance masks from the SAV dataset, our

Prompt 2: Static Instance Description

You are given two aligned images of the same scene:
1) The original image where the target instance is marked with a green outline (the outline is an overlay, not part of the object). 2) A version of the original image where the background is blurred to isolate the target instance.
Objective:
Write exactly one English description that refers only to the target instance and its scene/context in the original image.
Strict Rules:
* Focus: Describe only the focused object (the sharp item with green outline); ignore all other discrete objects.
* Content: If visible, include the object's visual attributes, function, or role, and its environmental context.
* Grammar: Do not use verbs or describe dynamic actions; use only nouns, adjectives, and prepositional phrases; structure around a single primary object noun phrase.
* People: Do not describe any person's actions or body poses.
* Exclusions: Do not mention the green outline, background blur, black paddings, image cropping, or other image artifacts; do not include quantities or numbers; do not describe the positional relationship between the camera and the target instance (e.g. from behind).
* Color Determination: Identify color only from the object's internal surfaces and textures, excluding any borders or edges affected by overlays. If the true color is uncertain, omit color adjectives.
* Outline Ignorance: Never describe the object as "green" or assign any color due to the green outline; the outline must be ignored for all attributes.
* Style: Objective, specific, and under 30 words.
* Validity: If any object other than the focused instance is described, the output is invalid.
Output Format:
* Exactly one sentence, under 30 words.
* Begin with a single generic category noun in brackets, followed by a colon and the description.
* Format: "[noun]: description."

Prompt 3: Dynamic Instance Description

You are an AI video analysis model specializing in tracking and describing the dynamics of a single object over time.
You receive two inputs:
1. **Text Description**: A sentence identifying the target object and its surrounding scene and context.
2. **Video Clip**: A silent video focused on the target object. The object is highlighted with a green border for tracking purposes.
Task: Generate a time-stamped log detailing the specific dynamics of the specified object shown in the video.
Rules:
- Source of Truth: The video clip is the source of truth. The text input is for context only. If a frame is completely black, it signifies the object is absent from the original footage during that time and do not create an entry describing the absence.
- Focus on Dynamics: Describe the object's actions, movements, transformations in state, interactions with others, and group activities in detail. Do not describe static properties if the object remains unchanged. If the object is motionless for the entire video, provide a single entry for the full duration.
- No Direction: Do not describe movement directions (e.g., left/right/up/down).
- Ignore Occlusion and Camera Motion Artifacts: Describe only changes inherent to the object itself or caused by direct interaction. Do not describe visual changes (such as fragmentation) caused by foreground occlusion and camera motion (e.g., the object being partially hidden by something passing in front of it or the object moves out of frame becuase of the camera motion).
- Camera vs Object Motion: Distinguish camera motion from object motion. Do not attribute camera movement to the object. Describe only the object's movement; do not describe camera movement.
- Ignore Visual Aids: Do not mention the green border lines and black paddings in the description.
- Consolidate Time: Merge continuous periods of the same action or inaction into a single time-stamped entry.
Strict Output Format (no extra commentary or sentences):
'[MM:SS] - [MM:SS]: [instance dynamics description]'

pipeline calculates the bounding box for the target instance in each frame and generates a new, smaller video. This clip is padded to a uniform size and dynamically re-centered in each frame to ensure the instance remains the focus of the shot. This isolation is crucial for the final stage of dynamic description, as it provides the VLM with an unambiguous view of the instance's actions, free from the interference of other objects or background events.

Stage 2: Static Description for Contextual Understanding To capture essential contextual information, we generate a static description based on the full, uncropped video frame. For each instance, we identify a single "best" representative frame from the original video, typically one where the object is clearly visible and centrally located.

We then present two images to the VLM: (1) the original, full-resolution frame with the target instance highlighted by a green outline, and (2) the same frame but with the background heavily blurred, further emphasizing the instance. This dual-image input allows the model to observe the object within its complete environment, facilitating a more holistic understanding. The VLM is then prompted to provide a concise, single-sentence description of the object's appearance and its role within the scene.

The prompt used for this stage is shown Prompt 2.

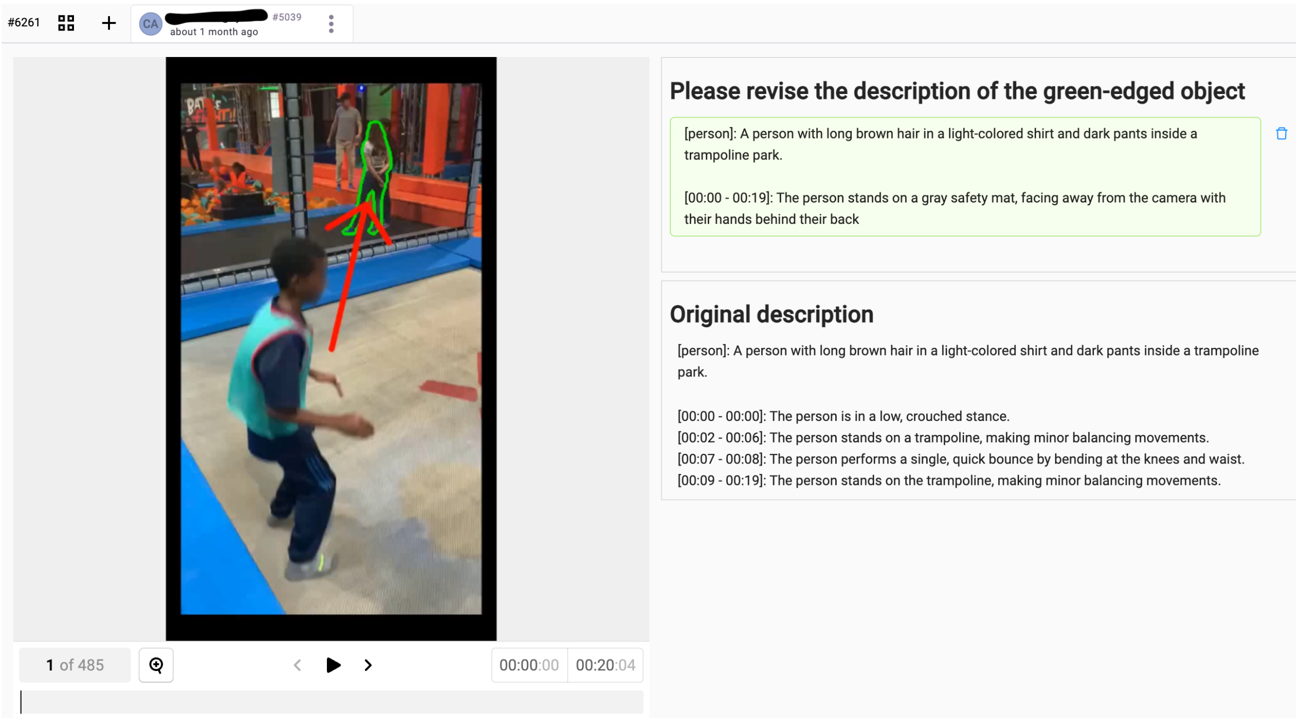


Figure 8. **Interface of Human Annotation for Instance Description (Sec. 10.3).** The left part displays the video, and on the right there is a text box where the annotator can freely edit or modify the content, based on original description that is AI-generated (Sec. 10.2).

Stage 3: Dynamic Description for Capturing Temporal Behavior With a static, contextual description in hand, the final stage is to capture the instance's dynamic behavior over time. For this, we use the cropped video clip generated in Stage 1. This focused input prevents the model from being distracted by other scene elements. We provide the VLM with both the cropped video and the static description generated in Stage 2. The static description acts as a crucial anchor, informing the model about *what* the object is, while the video shows *what the object does*. The model is prompted to produce a detailed, time-stamped log of the object's movements, actions, and changes in state, effectively creating a fine-grained temporal narrative for each instance. These detailed, dynamic descriptions form the basis from which our causal questions and evidence chains are generated.

The prompt used for this stage is shown in Prompt 3.

10.3. Human Annotation for Instance Description

The AI-generated captions from Sec. 10.2 provide dense coverage for every tracked instance, yet they remain vulnerable to the very issues CaST-Bench is designed to surface: cluttered scenes, tiny objects, and subtle causal cues often mislead frontier VLMs. To avoid propagating these hallucinations into later QA generation stages, we added a dedicated human verification pass that rewrites each instance description so that it is temporally faithful to the underlying video and linguistically precise.

Interface and Workflow We built a custom web interface (see Fig. 8) that streams the source video on the left panel while highlighting the target instance with the same green edge used throughout the pipeline. The right panel hosts an editable textbox seeded with the VLM-produced, time-stamped description plus controls for adding or removing timeline rows. Annotators proceed in two steps: (1) watch the full clip to understand the object's behavior, and (2) iteratively edit the draft description until every time span and action aligns with the video evidence. This tight coupling between visualization and editing allows annotators to immediately localize contradictions and keeps the overall ST formatting consistent with downstream tooling.

Annotation Guidelines We distilled the guideline into three concrete requirements. *(i) Contradiction removal*: every sentence must be cross-checked against the clip; any mismatch between the AI text and the video, including incorrect attributes, missed interactions, or spurious events, must be corrected before submission. *(ii) Temporal correctness*: if misaligned time breaks cause errors, annotators are instructed to split, merge, add, or delete timeline rows while preserving the canonical "[start–end]: action" format so that causal evidence extraction remains deterministic. *(iii) Special cases*: when the interface shows an "Absent" frame, annotators leave the corresponding time range empty; if the instance is too small or blurred to describe reliably, they either click "Skip" or enter "null" to flag the sample. These rules ensure that only verifiable, grounded content flows into the causal chain generator.

Annotator Expertise and Quality Control We recruited seven professional annotators who are native English speakers to maximize both factual accuracy and grammatical fluency. Each annotator underwent a calibration phase using the written handbook accompanying the interface, during which we sampled revisions for spot checks and provided feedback on missed contradictions. The combination of expert annotators, a purpose-built tool, and strict revision policies yields 10,728 timestamp-accurate instance descriptions that reliably serve as seed evidence for CaST-Bench.

10.4. QA and Causal Chain Generation with Explicit ST Evidences

Motivation and Challenges After the human-verified instance narratives in Sec. 10.3, the next bottleneck is converting thousands of time-stamped descriptions into causal QA items whose answers are only recoverable when a model reconstructs the same spatio-temporal chain. The scenes we curate often contain multiple overlapping human-object interactions, so naively prompting a VLM frequently yields generic "what" questions that ignore the causal cues we need. Furthermore, even when high-quality QA proposals emerge, they must be expressed in a structured format that later stages (distractor generation, mask-based validation, metric computation) can parse without manual post-processing. Finally, each textual evidence snippet must be tied back to the exact pixel tracks that human annotators produced; otherwise the benchmark would fail to measure grounded reasoning. The pipeline described below addresses these pain points in tightly coupled steps.

Step 1. Consolidating instance evidence into a causal reasoning canvas For every SAV video directory, we scan all verified caption JSONs, skip any file explicitly flagged as unusable, and extract the instance identifier together with its static appearance sentence and dynamic timeline. A normalization routine merges the static and dynamic sentences into a single "Combined Video Descriptions" block (as shown in Prompt 4) where each entry begins with the instance identifier followed by the appearance rationale and the chronological action log. This formatted block is the only textual context the downstream VLM sees; it forces the model to reason over causal candidates spanning different people, tools, and time ranges.

Step 2. Prompting the model as a causal QA author The generation step issues a single, large instruction to a video-aware language model. The instruction injects five ingredients: (i) the combined descriptions, (ii) the raw video, (iii) the causal question taxonomy (Prompt 6), (iv) a JSON schema (Prompt 5) that the output must obey. The model is required to return *exactly* six QA-Evidence items at once so that diversity is enforced within a single decoding call, and the string formatter verifies that every placeholder in the prompt has been replaced before the request is sent. The exact instructions are reproduced in Prompt 4).

Step 3. Enforcing a machine-checkable schema Even with a careful prompt, large models sometimes drift away from the desired structure. We therefore include the authoritative JSON layout in every call and validate the response immediately after decoding. Any malformed response is wrapped into an "error" JSON so that the dataset builders can retry the sample later. The schema is shown in Prompt 5; note how every field explicitly asks for temporal boundaries and instance identifiers, which later enable automatic alignment with bounding boxes.

Step 4. Covering the full causal taxonomy We embed the entire set of question types inside the prompt and instruct the model to cycle through subcategories when possible. This design choice directly supports the taxonomy introduced in the main paper (Sec. 4) and ensures that the resulting benchmark stresses explanation, counterfactual, prediction, and inference evenly. The taxonomy text shipped to the model is given at Prompt 6.

Step 5. Injecting explicit spatio-temporal evidence Once a valid JSON exists, a lightweight post-processor takes the question subject and every evidence instance, parses their reported start/end times in `mm:ss` format, and converts them into integer-second indices. It then loads the dense bounding-box tracks collected earlier (one JSON per tracked instance) and slices the corresponding coordinates for each second within the requested range. Two augmentations are written back into the QA structure: (i) a stable reference to the source track file ("SAV_instance_id") and (ii) a dictionary named "bboxes_in_range" that maps every aligned second to the $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ box recorded at 1 FPS.

10.5. Multiple-Choice Question Generation

To create a rigorous multiple-choice evaluation format, we developed a sophisticated, multi-stage pipeline to generate a set of plausible yet incorrect options (distractors). The primary challenge is to design distractors that probe for genuine causal reasoning, making it difficult for models to arrive at the correct answer by relying on superficial linguistic

Prompt 4: Causal QA Generation

You are a video analysis expert specializing in causal reasoning. Watch the video and read the time-stamped, instance-level descriptions, and then generate six high-quality causal QA-Evidence items. The output must strictly follow the provided JSON template.

INPUTS:
- Original video - Combined Video Descriptions (instance-level descriptions with timestamps) - Causal Questions Taxonomy - QA-Evidence Template (authoritative schema) - Example QA-Evidences

Combined Video Descriptions: {captions_text}

Causal Questions Taxonomy {question_types}

QA-Evidence Template {qa_evidence_template}

TASK:
Create six QA-Evidence following subcategories in the Causal Questions Taxonomy. Produce a single JSON object that exactly matches the QA-Evidence Template, with top-level key 'QA_Evidence_List' containing exactly six QA items.

FIELD SPECIFICATION (align with the QA-Evidence Template):
- Question - question_text: The question text. Do not include any instance_id. - subject_instance_id: if the question_text is asking about a specific subject, find the corresponding instance_id; otherwise null. - question_start_time & question_end_time: if the question_text mentions a specific time period, write "mm:ss"; otherwise null. - question_type_subcategory: one subcategory number from Causal Questions Taxonomy - question_type_subcategory_name: exact name of the subcategory - video_scene_category: Concise scene label to describe the scene. - Answer - answer_text: within 15 words. Do not include any instance_id. - Evidence - Include at least 1, at most 5 evidences. - Each evidence must include evidence_start_time, evidence_end_time, evidence_instance_id, and a concise evidence_rationale. - Distractor - Include exactly one distractor option. Do not include any instance_id. - Match the length and complexity of answer_text to avoid guess-by-length.

CONSTRAINTS:
- Question - Do not include any instance_id. Make the subject uniquely identifiable via appearance, actions, location, and timing descriptions. - Questions must have human as the subject; phrase the question so the human is the grammatical subject. Do not use non-human entities (objects, devices) as the subject in the question. - Avoid generic questions that could be answered by observing cues common to most instances in the scene; prefer questions that must depend on a small subset of specific evidences to be answered correctly. - About the small subset of specific evidences, prefer questions that require multiple different instances to answer, rather than those that can be answered using only a single instance. i.e., we prefer questions that should be answered by observing multiple evidences. - Questions must be answerable solely from the video; the test taker will not have access to the Combined Video Descriptions. - The question must be extremely difficult for an AI model to answer correctly without genuinely understanding the nuanced details of the video context. - Answer - Answer should be concise and uniquely correct. within 15 words. Do not include any instance_id. - Evidence - evidence_instance_id may refer to either a human or an object that supports the reasoning. - Make each evidence decisive and necessary: if that visual evidence were occluded or removed from the video, the question would become unanswerable or the answer would no longer be uniquely determinable. - Choose a minimal, non-redundant set of decisive evidences; avoid including evidence that is merely suggestive or duplicative. - Distractors - Ensure the distractor option is incorrect but plausible; mutually exclusive with the correct answer and contradict the video/descriptions. - You may create distractors basing on another salient instance in the video or commonsense expectations about the scene, such as inventing a cause or event that sounds logical or physically possible but did not happen in the video. - The distractor MUST mimic the linguistic structure, length, and level of detail of the Correct Answer, so that it is extremely difficult for an AI model to choose correctly. - General - Ground all content strictly in the provided descriptions/video; do not hallucinate unseen entities, times, or actions. - Do not alter, rename, or reassign any instance_id. Use instance IDs exactly as defined in the Combined Video Descriptions for both subject_instance_id and evidence_instance_id to ensure traceability and alignment.

OUTPUT REQUIREMENTS: - Single, valid JSON object that exactly matches the QA-Evidence Template. No prose before or after the JSON. No extra commentary or sentences.

FINAL VALIDATION STEP 1: Try to answer each question without watching the video: After generating the QA, please try to answer the question without watching the video. If you can answer correctly with only common sense reasoning but without any information from the video, please regenerate the QA because the original QA is too easy to answer.
FINAL VALIDATION STEP 2: Watch the Video to validate each generated QA-Evidence item. Confirm that: - The question subjects and evidence instances correspond to the correct pixel regions and entities observed in the video. - The causal relationships implied in the question and answer are visually justified — i.e., the cause and effect are temporally and spatially consistent. - The Answer is unique and unambiguous, directly supported by the cited evidences without conflicting viusal cues elsewhere. - The distractor option is clearly incorrect yet plausible, ensuring it could mislead someone who has not fully reasoned through the causal sequence. If any of these checks fail, regenerate or revise that QA-Evidence item until all conditions are fully satisfied.
FINAL VALIDATION STEP 3: (before returning): - JSON parses; no code fences; no placeholders remain. - question_text and answer_text contains no instance_id; subjects are uniquely identifiable from wording and timing. - All instance_ids exactly match the IDs in Combined Video Descriptions; none are renamed or re-assigned.

Prompt 5: QA_Evidence Template


```
{
    "QA_Evidence_List": [
      {
        "Question": {
          "question_text": "[Question text placeholder]",
          "subject_instance_id": "[Instance ID placeholder]",
          "question_start_time": "[Time placeholder]",
          "question_end_time": "[Time placeholder]",
          "question_type_subcategory": "[Number placeholder]",
          "question_type_subcategory_name": "[Category name placeholder]",
          "video_scene_category": "[Scene category placeholder]"
        },
        "Answer": {
          "answer_text": "[Answer text placeholder]"
        },
        "Evidences": [
          {
            "evidence_start_time": "[Time placeholder]",
            "evidence_end_time": "[Time placeholder]",
            "evidence_instance_id": "[Instance ID placeholder]",
            "evidence_rationale": "[Rationale text placeholder]"
          },
          (up to 5 evidence items like the first one)
        ],
        "Distractor": {
            "distractor_option_text": "[Distractor option text placeholder]"
        }
      },
      (totally 6 QA-Evidence items like the first one)
    ]
}
```


patterns or visual biases. Our approach is motivated by the need to mitigate the influence of confounders, as illustrated in Fig. 2, ensuring that a correct answer is selected based on a sound understanding of the causal chain of events in the video rather than spurious correlations. The process involves generating two main types of distractors, text-based and video-based, and then further refining them to create subtle variations that test the precision of a model's understanding. The generation process unfolds in three main steps:

Step 1: Text-Based Distractor Generation The first step is to create a distractor based only on the textual content of the question. Without access to the video, a large language model is prompted to generate a plausible alternative answer. The model is instructed to create an answer that is fundamentally different from the correct one, relying on common sense and the context provided in the question alone. The prompt emphasizes that the generated answer must not introduce new objects, people, or elements that are not mentioned in the question, and it must be concise (within 15 words) to match the style of the correct answer. This technique is designed to produce options that appeal to models with strong language priors but weak visual grounding, effectively testing for over-reliance on linguistic cues. If a model selects this distractor, it indicates that the model is relying on spurious linguistic correlations rather than genuine visual evidence from the video.

Step 2: Verification of Semantic Exclusivity After generating a text-based distractor, we ensure that it is semantically distinct from both the correct answer and the pre-existing video-based distractor (which is created from causally irrelevant objects in the video, as described in Step 3 of the QA generation process). Another language model is prompted to compare pairs of options and confirm that they do not carry the same meaning. The verification prompt asks the model to determine whether two answers have the same meaning, requiring a binary response. If a generated distractor is found to be a mere paraphrase or synonym of another option, it is discarded, and the generation process is repeated up to a maximum number of attempts. This verification step is crucial for maintaining a set of genuinely different choices, as semantically equivalent options would not provide meaningful discrimination between models with different levels of understanding.

Step 3: Near-Miss Option Generation To further increase the difficulty and test the precision of a model's understanding, we generate "near-miss" versions for each of the three options: the correct answer, the text-based distractor, and the video-based distractor. A language model is tasked with creating a new version of each option that adheres to the same sentence structure and template but alters exactly one key attribute. The prompt specifies that each near-miss option must match the length within a small tolerance (approximately 10% of the original) and that all six resulting options (three originals plus three near-misses) must be mutually exclusive. This creates a set of highly similar pairs of options, where only one is correct. This forces the model to perform fine-grained reasoning and pay close attention to critical details, as a superficial understanding would be insufficient to distinguish between the correct option and its near-miss counterpart. The near-miss design is particularly effective at testing whether models can identify subtle but decisive differences in causal reasoning chains.

Through this pipeline, each question is equipped with a comprehensive set of six options: the correct answer, a text-based distractor, a video-based distractor, and a near-miss version for each. This design enables a fine-grained analysis of model failures, distinguishing between errors caused by linguistic biases (selection of text-based distractors), visual misinterpretations (selection of video-based distractors), or a lack of detailed causal understanding (selection of near-miss options). The multi-layered distractor strategy ensures that achieving high accuracy on CaST-Bench requires models to construct precise, grounded causal chains rather than relying on surface-level cues.

10.6. Mask-Based QA Filtering for Causal Chain Validation

After generating QA pairs with causal chains, we must ensure that questions genuinely require the visual evidence we provide and cannot be answered through spurious correlations or incomplete reasoning chains. As discussed in Sec. 4, mitigating the influence of confounders is fundamental to causal inference, and our filtering procedure addresses this by validating both the necessity of visual evidence and the completeness of the causal chain. We employ a three-step filtering process: first, a text-based filter removes questions answerable without video access; second, a novel video-masked filter validates that questions require the specific spatio-temporal evidence we provide; and finally, human annotators conduct thorough review to ensure quality.

Step 1: Text-Based Filtering Some questions may be answerable through linguistic patterns alone without requiring visual evidence, allowing models to rely on spurious linguistic correlations rather than genuine visual understanding. To ensure questions require actual video evidence, we

Prompt 6: Causal Question Taxonomy

```
## CATEGORY 1: CAUSAL EXPLANATION
**Purpose:** Questions that explain the reasons (why) or mechanisms (how) behind actions or events,
based on spatiotemporal evidence.

### Subcategory 1.1: Why Questions (Causal Reasons)
**Description:** These questions seek to understand the underlying causes and motivations behind ob-
served actions or events. They require reasoning about causal chains and identifying the root causes
that led to specific outcomes.
**Causes may include:**
- Direct triggers (immediate causes that set events in motion)
- Enabling conditions (circumstances that made the action possible)
- Mediating mechanisms (intermediate processes that connect cause to effect)
- Rule-based causality (social, legal, or procedural reasons)

### Subcategory 1.2: How Questions (Mechanisms)
**Description:** These questions focus on understanding the processes, methods, and mechanisms throu-
gh which specific outcomes were achieved. They require analyzing the step-by-step procedures and tec-
hniques used.
**Mechanisms may include:**
- Sequential steps (ordered procedures and workflows)
- Tool use (utilization of instruments and equipment)
- Exploitation of constraints/supports (leveraging environmental or structural elements)
- Path/obstacle handling (navigation and problem-solving strategies)
- Socially/rule-driven processes (following protocols and social conventions)

## CATEGORY 2: COUNTERFACTUAL REASONING
**Purpose:** Questions that test the ability to deduce the most direct and unavoidable consequence of
a single, precisely defined alteration to a key causal element within the scene, while holding all o-
ther scene variables constant.
**Mechanism for Uniqueness**: The question must isolate a single variable and propose a minimal, con-
crete change. The correct answer is the primary effect that follows directly from this change accord-
ing to physical laws, established procedures, or strong social norms, before any secondary reactions
or adaptations can occur.

### Subcategory 2.1: Physical Counterfactual
**Description:** Questions that explore the immediate physical outcome after altering a single mater-
ial property, environmental condition, or piece of equipment. The change must be specific enough to
trigger a deterministic outcome based on physics (e.g., gravity, momentum, structural failure).
```

Causal Question Taxonomy (Continued)

- Equipment failure (malfunction of tools, devices, or systems)
- Environmental alteration (changes in weather, lighting, terrain, or surroundings)
- Removal of supports/constraints (elimination of structural or physical aids)
- Parameter changes (modifications to speed, strength, size, or other physical properties)
- Alternative actions (different physical approaches or methods)

Subcategory 2.2: Social Counterfactual
Description: Questions that explore the immediate social outcome after altering a single, critical social variable|such as a specific instruction, the removal of a person with a unique role, or a violation of a clear protocol. The outcome should be the most direct consequence based on the established social or rule-based dynamics.
- Absence of participants (removal of key people from the scenario)
- Behavioral shifts (changes in individual or group behavior patterns)
- Altered instructions or rules (modifications to guidelines, protocols, or social norms)
- Communication differences (changes in how people interact and exchange information)
- Group dynamics changes (shifts in team composition, leadership, or social hierarchy)

CATEGORY 3: PREDICTIVE ANTICIPATION
Purpose: To generate questions that require predicting the most probable and immediate outcome or next action in a sequence, based on a clearly established trajectory (physical, intentional, or procedural) derived from unambiguous spatiotemporal evidence.
Mechanism for Uniqueness: The question must be based on a scene where actions are already in motion or a clear preparatory state has been established. The prediction should be limited to the logical continuation or completion of this ongoing action within the next moment (e.g., 1-2 seconds), thereby excluding more distant or speculative possibilities.

Subcategory 3.1: Physical Prediction
Description: Questions that predict the immediate future state of an object or system based on its current motion, trajectory, and interaction with the immediate environment. The scenario must provide sufficient physical information (e.g., velocity, direction, position relative to obstacles) to make the next state highly deterministic.
Focus: Predicts outcomes based on physical laws:
- Motion trajectory (path and velocity of moving objects)
- Stability (balance, structural integrity, and equilibrium states)
- Collisions (interactions between objects and their consequences)
- Environmental effects (impact of weather, terrain, and external conditions)
- Threshold states (critical points where systems change behavior)

Subcategory 3.2: Behavioral Prediction
Description: Questions that predict a person's next immediate action based on a sequence of behavior that clearly indicates a short-term goal or intention. The evidence must include preparatory actions, focus of attention (gaze), and body posture that collectively point to a single, imminent action.
Focus: Predicts human actions based on:
- Behavioral patterns (consistent ways individuals or groups typically act)
- Habits (routine behaviors and automatic responses)
- Social norms (expected behaviors within cultural or social contexts)
- Immediate intentions (short-term goals and immediate objectives)
- Interaction cues (non-verbal signals and social indicators)
Note: This subcategory inherently focuses on humans; ensure animals/human-operated objects are included only if tied to the actor's behavior.

CATEGORY 4: INFERENTIAL DESCRIPTION
Purpose: Questions that infer implicit attributes or states that are not directly visible or explicitly stated, grounded in multiple observable cues (not mere description).

Subcategory 4.1: Identity/Role Inference
Description: These questions require deducing a person's professional role, social status, or identity based on observable indicators rather than explicit information. They involve pattern recognition and social knowledge.
Description: Infers a person's role or profession from:
- Clothing (uniforms, professional attire, protective gear, or status indicators)
- Tools (specialized equipment, instruments, or devices characteristic of specific roles)
- Directive behavior (leadership actions, supervisory activities, or authoritative gestures)
- Others' responses (following instructions, seeking help, coordinating around the person)

Subcategory 4.2: Interpersonal Relationship Inference
Description: Determine the relationship type between people (e.g., colleagues, supervisor-subordinate, caregiving/affectional) based on interaction style, proximity, and collaboration patterns.
Evidence cues:
- Distance and touch (intimate contact, protective stance, personal space)
- Collaboration and division of labor (signs of ongoing coordination, complementary tasks)
- Communication style (formal vs. informal, directive vs. consultative)
- Context and setting (home/work/public-service role expectations)

Causal Question Taxonomy (Continued)

Subcategory 4.3: Goal/Intent Inference (non-immediate)
Description: Infer a person's short-term goal or task state without predicting the next second of action (to avoid overlap with predictive categories).
Evidence cues:
- Preparation and planning (arranging/checking tools, tuning equipment, consulting a list)
- Attention and path (gaze target, heading, pointing gestures)
- Rule/procedure alignment (standard layouts, adherence to task steps)

Subcategory 4.4: Emotion/Motivation Inference
Description: Infer likely emotional state or underlying motivation (e.g., tense, relaxed, urgent, helpful) from facial cues, action tempo, and situational context.
Evidence cues:
- Face and eyes (muscle tension, avoidance vs. steady gaze)
- Body tempo (hurried vs. slow, hesitation, force modulation)
- Situational events (setbacks, threats, recent completion of a goal)

Subcategory 4.5: Skill/Expertise Inference
Description: These questions involve assessing a person's level of competence, expertise, or professional skill based on their performance quality and behavioral indicators. They require understanding of skill development and professional standards.
Focus: Infers level of skill or professionalism from:
- Fluency (smoothness and efficiency of task execution)
- Correctness (accuracy and adherence to proper procedures)
- Error rates (frequency and types of mistakes made)
- Precision in performance (attention to detail and quality of work)
- Confidence indicators (certainty, hesitation, and self-assurance)
- Others' responses (seeking help vs. giving guidance)

Subcategory 4.6: Ownership/Belonging Inference
Description: These questions involve determining who owns or has rights to specific objects, spaces, or resources based on behavioral patterns and spatial relationships. They require understanding of ownership indicators and territorial behavior.
Focus: Infers ownership of objects based on:
- Placement (where objects are positioned relative to people)
- Guarding (protective behaviors and territorial actions)
- Frequency of use (how often and how naturally someone interacts with objects)
- Access patterns (who has permission to use or modify items)
- Personalization (customization and individual modifications)

Subcategory 4.7: Risk/Safety Assessment
Description: These questions involve evaluating potential dangers, safety levels, and risk factors in a given situation. They require understanding of safety indicators, hazard recognition, and risk assessment principles.

Focus: Infers level of hazard from:
- Unstable structures (structural integrity and potential collapse points, heat, sharp edges, fall risk, load issues)
- Warning signs (safety indicators, barriers, and cautionary signals)
- Crowd and individual reactions (avoidance, alarm, cautious handling)
- Environmental conditions (weather, lighting, terrain, and external threats)
- Safety equipment (presence or absence of protective gear and safety measures)

Subcategory 4.8: Success/Failure Inference
Description: These questions involve determining whether an action, task, or process is succeeding or failing based on observable outcomes, behavioral reactions, and performance indicators. They require understanding of success markers and failure signals.
Focus: Infers success or failure from:
- Outcome indicators (visible results and achievement markers)
- Behavioral reactions (satisfaction, frustration, celebration, or disappointment)
- Performance quality (accuracy, efficiency, and adherence to standards)

Subcategory 4.9: Environmental Condition Inference
Description: Infer weather/temperature/surface/lighting or similar environmental conditions from human adaptations, equipment, and environmental traces.
Evidence cues:
- Adaptive behavior (seeking cover, tightening clothing, squinting, slow gait)
- Equipment usage (umbrellas, flashlights, hand warmers, crampons)
- Surface traces (puddles, reflections, dust, ice)
- Lighting and visibility (glare, backlight, dimness, fog)

Subcategory 4.0: Other Contextual Inference
Description: Infer factors not covered above but relevant to understanding the scene, such as time of day, culture/customs, economic/resources, or rules/regulations.
Evidence cues:
- Time cues (shadow length, store open/closed state, festive decorations)
- Cultural/customary markers (attire styles, rituals, language/iconography)
- Economic/resource signs (equipment age, material sufficiency)
- Rules/regulation (signage, procedures, checkpoints)

present the question and all answer options to a language model without providing the video. If the model successfully identifies the correct answer, this indicates the question can be solved through text-only reasoning and is filtered out.

Step 2: Video-Masked Filtering Even when questions pass the text-based filter, they may still be answerable through unintended visual cues, dataset biases, or incomplete causal chains that miss critical evidence. The challenge here is to validate that the causal chain we provide is complete — that is, *questions should truly require the specific spatio-temporal evidence we identify, and removing this evidence should make the question unanswerable*. This validation is motivated by the need to ensure that our benchmark tests genuine causal reasoning rather than superficial pattern matching or reliance on unintended visual context.

To address this, we implement a video-masked filtering procedure. First, we collect all bounding box regions associated with both the question subject (if specified) and all evidence instances from the causal chain. These regions represent the spatio-temporal evidence that should be necessary to answer the question. We then create a masked version of the video where these evidence regions are blacked out by setting the corresponding pixels to black.

The masked video is then presented along with the question and all answer options to a vision-language model. The model is instructed to answer based only on visible evidence from the video and to select a special option indicating "Unable to determine" if insufficient information is available to make a confident choice. If the model can still correctly answer the question using the masked video, this indicates one of three problematic scenarios: (1) the causal chain is incomplete, missing critical evidence that would be necessary for a human to answer correctly; (2) the question can be solved through unintended visual cues outside the masked regions, such as background context or other objects that happen to be visible; or (3) dataset biases or question wording allow correct answers without the intended evidence. Questions that pass this masked-video test are filtered out, ensuring that only questions requiring the complete, intended causal chain remain in the benchmark.

After these automated filtering steps, human annotators conduct a final review and revise to validate the question, answer, and causal chain, ensuring that the remaining questions meet our quality standards. Through this rigorous three-step filtering process, only approximately 40% of the initially generated QA pairs remain, as reported in Sec. 4, resulting in a benchmark that truly tests causal reasoning grounded in spatio-temporal evidence.

11. Details of Experiment Setup

11.1. Evaluation Prompt

All evaluated VLMs shared a single unified prompt for video QA. For the multiple-choice setting, the exact prompt is provided in Prompt 7.

11.2. VLM Hyperparameter Configuration

We configure all VLMs with `max_new_tokens` set to 2048, limiting each sample to at most 2,048 generated tokens (excluding the input length), and `cutoff_len` set to 204,800, which caps the combined input-plus-output length at 204,800 tokens so that truncation is effectively disabled and long contexts are preserved end-to-end.

12. Evaluation Suite

12.1. Grounded Causal Chain Evaluation

Evaluating the correctness of a predicted causal chain is fundamentally harder than checking a single grounding target. A model must recover every actor that participates in the causal process, align their evidences across different time ranges, and ensure the supporting boxes stay faithful to the underlying video. Because the number and ordering of predicted instances rarely match the ground truth, we design a three-stage procedure: (i) greedy instance matching, (ii) temporal accuracy scoring, and (iii) spatio-temporal accuracy scoring. The latter two stages use the IM-tIoU and IM-vIoU metrics defined in Sec. 4.3 in the main paper, so we expand here on the instance matching procedure that enables those metrics.

Instance Matching The central difficulty lies in linking each predicted instance to a unique ground-truth counterpart without relying on shared identifiers. A naive name-based match fails because models may rename entities, merge multiple actors, or hallucinate new ones. Instead, we compare every predicted–ground-truth pair using a spatio-temporal overlap score that multiplies the temporal IoU between their evidence spans and the average spatial IoU of the aligned bounding boxes over the overlapping frames. This score naturally favors predictions that are both time-aligned and spatially localized. Once we obtain the score matrix between all predicted instances and GT instances, we apply a greedy selection process: the pair with the highest score is matched first, both elements are removed from further consideration, and the process repeats until no positive-overlap pairs remain. This one-to-one assignment ensures that each ground-truth instance contributes at most one match, preventing duplicated credit for the same evidence and allowing downstream metrics to treat unmatched predictions as false positives and missing instances as false negatives.

Prompt 7: CaST-Bench Evaluation Prompt

```
<video> {question}

Please analyze the video and answer the multiple-choice question by selecting the most appropriate option, and
provide the visual evidences from the video that support your answer.

Your response should be in the following JSON format (Do not include ellipses; Ensure the JSON is self-contained
and valid):

OUTPUT TEMPLATE (replace with your content; do not include placeholders):
{
  "instances": [
    {
      "instance_name": "short name of the instance",
      "evidences": [
        {
          "evidence_start_time": "mm:ss",
          "evidence_end_time": "mm:ss",
          "evidence_rationale": "evidence description",
          "bboxes_in_time_range": {
            "ss": "[x_min, y_min, x_max, y_max]",
            "ss": "[x_min, y_min, x_max, y_max]",
            "ss": "[x_min, y_min, x_max, y_max]",
            ...
          }
        },
        ...
      ]
    },
    ...
  ],
  "answer_choice": "the chosen option"
}

STRICT OUTPUT ONLY
- Output a single valid JSON object. No prose, no code fences, no comments.
- Use double quotes for all strings. No trailing commas.
- Keep key names exactly as specified. Do not add, remove, or rename keys.
- Keep the key order as shown in the template.

SCHEMA AND CONSTRAINTS
- "instances": array of one or more instances, each with:
  - "instance_name": short noun phrase identifying the entity (string).
  - "evidences": array of one or more evidences for that instance.
- Each evidence must have:
  - "evidence_start_time": start timestamp in "mm:ss" (zero-padded).
  - "evidence_end_time": end timestamp in "mm:ss" (>= start, zero-padded).
  - "evidence_rationale": one concise sentence explaining this evidence clip.
  - "bboxes_in_time_range": object whose keys are every whole-second timestamp from "evidence_start_time" to
  "evidence_end_time" (inclusive) in "mm:ss" format (zero-padded), and whose values are strings formatted exactly
  as "[x_min, y_min, x_max, y_max]" with integers. No missing seconds.
- The sum of all evidences across all instances must be <= 5.
- "answer_choice": the single uppercase option letter that best matches the answer.
```

Temporal Accuracy (IM-tIoU) With the instance pairs fixed, we measure how accurately the model captured the timing of each evidence segment using the IM-tIoU metric introduced in Sec. 4.3. The metric averages temporal IoU scores across ground-truth instances, so any unmatched or poorly aligned instance directly lowers the final score.

Spatio-Temporal Accuracy (IM-vIoU) Finally, we apply the IM-vIoU metric from Sec. 4.3 to assess joint spatial and temporal fidelity. Because the greedy matching step guarantees clean one-to-one assignments, IM-vIoU can focus purely on how well each matched pair overlaps in both time and space, rewarding predictions that retrieve the correct actors, durations, and bounding boxes across the entire causal chain.

12.2. Open-Ended Evaluation

The quantitative metrics in Sec. 4.3 diagnose how well a model grounds causal evidence and selects multiple-choice answers, yet they cannot fully capture the nuanced, free-form responses that arise when a system explains its reasoning without option prompts. To complement the automatic metrics, we introduce an open-ended evaluation protocol that enlists a large language model (LLM) as an expert judge. This protocol focuses on two challenges highlighted throughout CaST-Bench: (i) ensuring that causal chains re-

Prompt 8: Open-Ended Evaluation

```
You are an expert evaluator of causal reasoning quality.

Given:
1.  A causal question about a video event {question}
2.  The ground-truth conclusion answer {gt_answer}
3.  The ground-truth causal reasoning process {gt_reasoning}
4.  A test-taker model's generated conclusion answer {pred_answer}
5.  A test-taker model's generated causal reasoning process {pred_reasoning}

Task:
Evaluate the test-taker model's reasoning across four dimensions. Assign
a separate score (0{10) for each dimension according to the standards below.

Evaluation Dimensions:
1. Answer Conclusion Correctness
* Compare only the model's conclusion answer to the ground-truth conclusion
  answer; ignore any reasoning content.
* Judge semantic equivalence, polarity, entity/attribute correctness, and
  numeric/unit consistency; penalize contradictions, material vagueness,
  or hedging that alters commitment.

2. Causal Chain Logical Consistency
* Evaluate only the generated answer and reasoning; do not reference ground-truth
  answer or ground-truth reasoning.
* Verify that causes precede effects and that the causal sequence is minimal yet
  sufficient to explain the answer.
* Identify any logical leaps, post-hoc reasoning, or teleological claims lacking
  justification.
* Ensure temporal feasibility and internal consistency across causal steps.

3. Evidence Coverage & Completeness
* Use the ground-truth causal reasoning as the reference. Check that the generated
  causal reasoning includes and aligns with its key entities, events, moments,
  and causal steps.
* Evaluate recall of essential causal steps and contextual conditions relative
  to the ground truth.
* Penalize missing core components, contradictions to the ground truth, or
  hallucinated steps; avoid rewarding overemphasis on partial evidence.

4. Evidence{Conclusion Overall Justification
* Consider the generated answer and the generated reasoning together: does the
  provided reasoning justify the stated answer, and do both align with the ground
  truth overall?
* Assess the logical coherence from evidence to conclusion, calibration of
  confidence, and global plausibility.

Scoring Standards for each dimension:
9~10: Exemplary performance with no flaws
7~8: Non-critical deviations present
5~6: Quality-impairing defects
3~4: Serious validity-compromising errors
0~2: Fundamental functionality failure

Output Format:
Output only a valid JSON object in the following format (no additional text):
{{"answer_conclusion_correctness": 00.00,
  "causal_chain_logical_consistency": 00.00,
  "evidence_coverage_completeness": 00.00,
  "evidence_conclusion_overall_justification": 00.00}}
```

main faithful to the video even when they are written in natural language, and (ii) rewarding models whose final answers are logically supported by their own rationales rather than by dataset priors.

We prompt an LLM to act as an impartial evaluator. The full instruction provided to the judge is reproduced at Prompt 8.

13. Case Studies and Failure Analysis

Beyond the quantitative ablation studies and error analysis presented in the main paper, we here perform a qualitative analysis of the case studies and failure patterns exhibited by the evaluated models.

Vulnerability to Spurious Visual Confounders A core design principle of CaST-Bench is the inclusion of distractors that target spurious correlations — visual or linguistic cues that are salient but causally irrelevant. We observe that even top-tier proprietary models can be misled by these confounders when they fail to strictly follow the causal chain. As shown in the failure example of Fig. 16, when asked *"How do customers know where to line up?"*, Gemini-2.5-Pro [9] correctly identifies visual elements in the scene (blue and yellow arrows on the floor). However, it fails to distinguish between *descriptive existence* (the arrows exist) and *causal relevance* (the arrows indicate product sections, not queue lines). The model falls into the trap of the "Video-Based Distractor" (Option A), ignoring the true causal evidence — the social cue of other customers already queuing (Option C). This error demonstrates that strong perception capabilities alone are insufficient; models must be able to filter out salient but non-causal visual information to reason correctly.

Visual Hallucination and Grounding Disconnect Another significant failure pattern is the generation of plausible-sounding textual rationales that are completely disconnected from the actual visual data — a phenomenon we term "grounding disconnect". In Fig. 17 (Failure Example #2), InternVL-3.5 [34] attempts to answer a counterfactual question about a traffic scene. While the model selects a distractor answer ("hit by the descending barrier"), it attempts to justify this choice by hallucinating evidence. The generated bounding boxes for Evidence #1 and #2 are incorrect and repetitive, and the rationale incorrectly states that the barrier is moving downward toward the person, even though the video shows no such movement at those timestamps. This means a model may produce a logically coherent textual explanation that is physically grounded in hallucinated pixels, proving the necessity of enforcing explicit ST evidence grounding.

Instruction Following and Formatting Failures The CaST-Bench task requires models to output a structured JSON containing precise timestamps and frame-by-frame bounding boxes. This poses a heavy instruction-following challenge. As illustrated in Fig. 18 (Failure Example #3), InternVL-2.5 [6] successfully identifies the context of the question but fails to generate valid coordinates. Instead of calculating the specific bounding box values, the model simply repeats the placeholder templates (e.g., $[x_{min_1}, \ldots]$) provided in the prompt. This "format collapse" renders the output invalid for evaluation. This failure pattern suggests that complex spatio-temporal reasoning tasks require not only visual understanding but also robust instruction-following capabilities to map internal reasoning into precise, structured outputs.

14. Social Impact, License, and Access

14.1. Broader Impact

CaST-Bench advances the field of VLMs by shifting the focus from surface-level perception to deep, grounded causal reasoning, a capability essential for sophisticated video analysis and anticipation tasks. By mandating that models validate their answers with explicit spatio-temporal evidence, the benchmark significantly enhances transparency and trust, enabling the distinct identification of genuine understanding versus reliance on spurious correlations or hallucinations. Furthermore, the work introduces a scalable Human-AI collaborative pipeline for constructing high-quality, dense annotations, offering a methodological blueprint for future complex reasoning dataset creation. Ultimately, CaST-Bench serves as a rigorous diagnostic tool to mitigate visual confounders and biases, fostering the development of more robust, interpretable, and reliable multimodal AI systems.

14.2. Limitations

Although our human–AI collaborative pipeline yields high-quality spatio-temporal causal annotations, it limits scalability. Extending CaST-Bench to additional domains or substantially increasing its size would require retracing multiple stages of human verification to maintain quality. The benchmark inherits biases from the SegmentAnything-Video (SAV) corpus [29], such as the geographic distribution and activity types favored in that source dataset. Moreover, despite detailed annotation guidelines, human editors may introduce subtle preferences when rewriting rationales or validating causal chains, which could affect linguistic style or emphasis.

14.3. Ethics, License, and Data Access

All videos originate from the publicly released SAV dataset, and we follow the source licensing terms when curating clips. During data construction we filter out samples that contain sensitive or unsafe content and require annotators to follow institutional ethical standards, including anonymizing identifiable details beyond what is already visible in the original footage. We will release CaST-Bench (data, annotations, and evaluation code) under the Creative Commons Attribution 4.0 International (CC BY 4.0) license and the MIT License. This license grants broad research use provided that derived works cite the benchmark, while the underlying SAV assets remain governed by their original licenses and terms of use. Large multimodal models (e.g., Gemini-2.5-Pro) are used within the pipeline to bootstrap descriptions and candidate QA pairs; every machine-generated artifact is subsequently reviewed and revised by trained annotators to guarantee factual alignment.

Question Type: **Causal Explanation - Why questions (reasons)**

Question: **Why does the child on the green balance bike stop moving between 00:01 and 00:08?**

Answer | Text-Based Distractor | Video-Based Distractor
Near-miss A. | Near-Miss T.B.D. | Near-Miss V.B.D.

"A": "Because the person in the long dark coat told the child to stop moving.",

"B": "Because the person in the long dark coat is adjusting the bike's handlebar.",

"C": "Because the person in the black puffer jacket is adjusting the child's helmet.",

"D": "Because the person in the black puffer jacket is adjusting the child's shoe.",

"E": "The person in the puffer jacket is holding the bike's handlebar, preventing it from moving.",

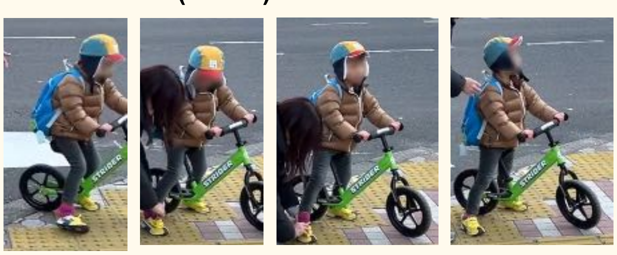
"F": "The person in the puffer jacket is holding the bike's seat, preventing it from moving."

Evidence #1 (cause):

Start Time: 00:01
End Time: 00:08

Rationale:
The person bends over and crouches next to the child, interacting with their shoe.

Evidence #2 (effect):



Start Time: 00:01
End Time: 00:08

Rationale:
The child remains stationary on the balance bike during this entire period.

Figure 9. **CaST-Bench Data Sample 1.** Question Type: Causal Explanation - Why questions (reasons).

Question Type: **Causal Explanation - How questions (mechanisms)**

Question: **How does the yellow ride-on toy move forward through the hallway?**

Answer	Text-Based Distractor	Video-Based Distractor
Near-miss A.	Near-Miss T.B.D.	Near-Miss V.B.D.

"A": “The child sitting on it is using their hands to pull it.",

"B": “The person in the magenta vest is pulling it from the front.",

"C": "It is a self-propelled toy that was activated off-screen.",

"D": “The child sitting on it is using their feet to push it.",

"E": “The person in the magenta vest is pushing it from behind.",

"F": "It is a remote-controlled toy being operated off-screen."

Evidence #1 (cause):

Start Time: 00:01
End Time: 00:10

Rationale:
This person is standing directly behind the toy and is walking forward, causing the toy to move.

Evidence #2 (effect):

Start Time: 00:01
End Time: 00:10

Rationale:
The child on the toy scooter is being pushed and is not propelling the toy themselves.

Figure 10. **CaST-Bench Data Sample 2.** Question Type: Causal Explanation - How questions (mechanisms).

Question Type: **Counterfactual Reasoning - Physical counterfactual**

Question: **If the person in the black puffy jacket had not looked ahead, what would have been the most direct physical consequence?**

Answer | Text-Based Distractor | Video-Based Distractor
Near-miss A. | Near-Miss T.B.D. | Near-Miss V.B.D.

"A": "They would have been struck by the approaching silver car.",

"B": "They would have been struck by the descending red and white barrier arm.",

"C": "They would have walked directly into the path of the silver car.",

"D": "The person would have been hit by the descending red and white barrier arm.",

"E": "The person would have been hit by the silver car.",

"F": "They would have walked into the path of the moving car."

Evidence #1 (cause):

Start Time: 00:01
End Time: 00:06

Rationale:
The person was initially walking on a trajectory that intersected with the moving car's path.

Evidence #2 (cause):

Start Time: 00:06
End Time: 00:08

Rationale:
The person stops their forward movement just as the car occupies their intended path.

Evidence #3 (cause):

Start Time: 00:01
End Time: 00:05

Rationale:
The silver car was moving forward but not in the direction the person was heading towards.

Figure 11. **CaST-Bench Data Sample 3.** Question Type: Counterfactual Reasoning - Physical counterfactual.

Question Type: **Counterfactual Reasoning - Social counterfactual**

Question: **What would happen if the toddler in the white patterned shirt suddenly ran away?**

Answer | Text-Based Distractor | Video-Based Distractor
Near-miss A. | Near-Miss T.B.D. | Near-Miss V.B.D.

"A": "The child could lose balance and fall.",

"B": "The man in the black t-shirt would stop attending to the stroller and pursue the toddler.",

"C": "The woman in the pink shirt would call out the toddler's name to get them to stop.",

"D": "The man in the black t-shirt would stop pushing the stroller and look for the toddler.",

"E": "The child could trip over the stroller and fall.",

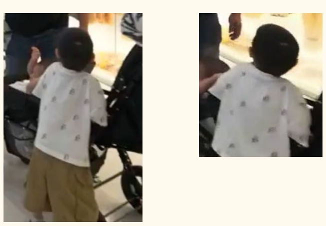
"F": "The woman in the pink shirt would stop pushing the stroller and pursue the toddler."

Evidence #1 (cause):

Start Time: 00:00
End Time: 00:01

Rationale:
The man is bent over, in close proximity to the toddler and stroller, indicating a supervisory role.

Evidence #2 (cause):



Start Time: 00:00
End Time: 00:01

Rationale:
The toddler is standing right next to the man, establishing a clear social unit of caregiver and child.

Figure 12. **CaST-Bench Data Sample 4.** Question Type: Counterfactual Reasoning - Physical counterfactual.

Question Type: **Predictive Anticipation - Behavioral anticipation**

Question: **Based on the actions of the person in the light blue-colored top, what is the most likely immediate next action they will perform after 00:10?**

Answer | Text-Based Distractor | Video-Based Distractor
Near-miss A. | Near-Miss T.B.D. | Near-Miss V.B.D.

"A": "They will turn back around and pick up something from the chair.",

"B": "They will pull the chair further out from the table.",

"C": "They will push the chair back under the table.",

"D": "They will turn back around and sit down in the chair again.",

"E": "They will walk away from the chair and towards the window.",

"F": "They will walk away from the chair and towards the exit."

Evidence #1 (cause):


Start Time: 00:05
End Time: 00:07

Rationale:
The person stands up and orients their body away from the table.

Evidence #2 (cause):


Start Time: 00:08
End Time: 00:10

Rationale:
The person puts on their backpack and begins to take a step, indicating departure.

Figure 13. **CaST-Bench Data Sample 5.** Question Type: Counterfactual Reasoning - Social counterfactual.

Question Type: **Inferential Description - Skill/expertise inference**

Question: **What is the most likely level of expertise of the person in the black shirt?**

Answer | Text-Based Distractor | Video-Based Distractor
Near-miss A. | Near-Miss T.B.D. | Near-Miss V.B.D.

"A": "He is a professional learning how to use new equipment.", (Near-Miss V.B.D.)

"B": "He is likely a director, producer, or a junior technician.", (Near-Miss T.B.D.)

"C": "He is a skilled professional working independently.", (Near-miss A.)

"D": "He is an amateur learning how to use the equipment.", (Video-Based Distractor)

"E": "He is likely a director, producer, or a senior technician.", (Text-Based Distractor)

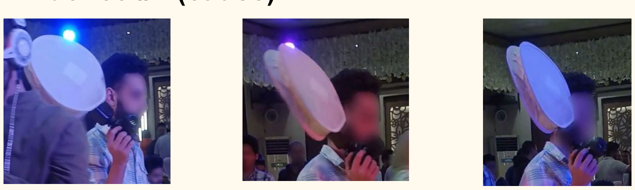
"F": "He is a skilled professional working as part of a crew." (Answer)

Evidence #1 (cause):

Start Time: 00:00
End Time: 00:14

Rationale:
He is operating a large, professional-grade camcorder, not a consumer device.

Evidence #2 (cause):



Start Time: 00:10
End Time: 00:14

Rationale:
Another person joins him with a light diffuser, indicating a coordinated, professional video crew.

Figure 14. **CaST-Bench Data Sample 6.** Question Type: Predictive Anticipation - Behavioral anticipation.

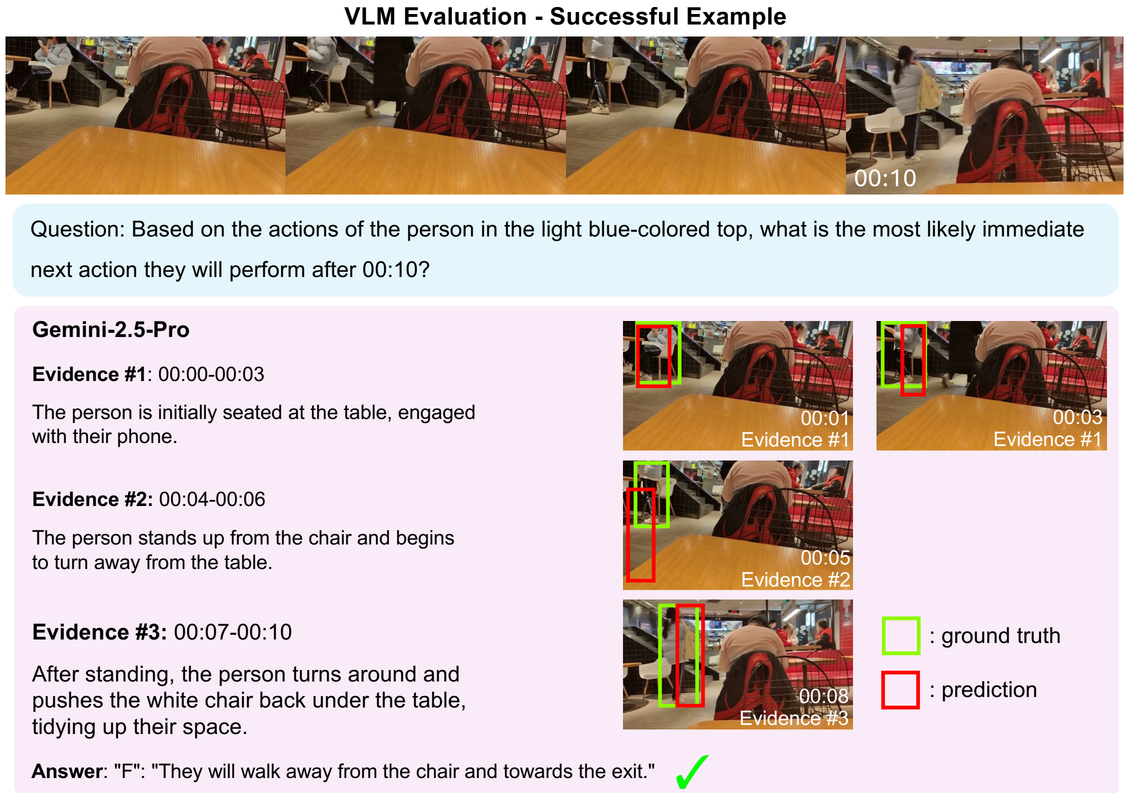


Figure 15. **Case Studies and Failure Analysis.** Successful example.

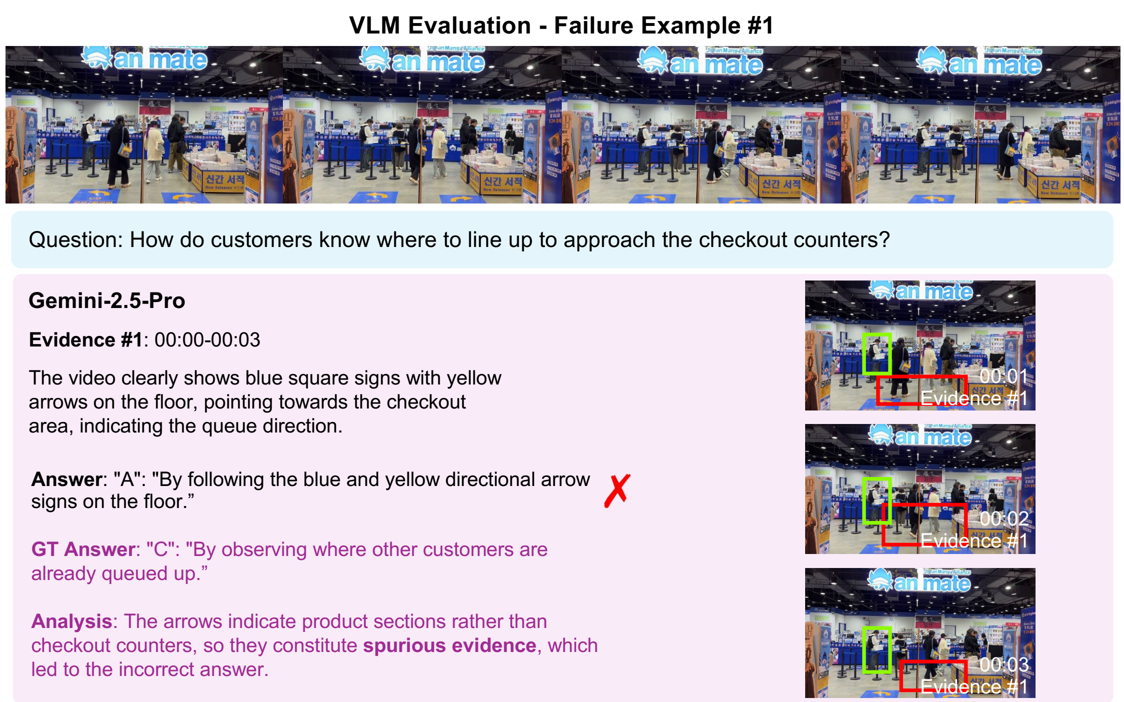


Figure 16. **Case Studies and Failure Analysis.** #1 failure example.

VLM Evaluation - Failure Example #2

Question: If the person in the black puffy jacket had not looked ahead, what would have been the most direct physical consequence?

InternVL-3.5

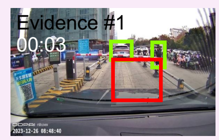
Evidence #1: 00:00-00:05
The person in the black puffy jacket is standing in the path of the silver car.

Evidence #2: 00:05-00:10
The red and white barrier arm is descending towards the person.

Answer: "D": "The person would have been hit by the descending red and white barrier arm." ✗

GT Answer: "F": "They would have walked into the path of the moving car."

Analysis: Although the textual causal chain appears reasonable, it does not match the video content and cannot reasonably ground the key evidence in the video.



Incorrect bounding boxes with repeated identical numbers.

Figure 17. **Case Studies and Failure Analysis.** #2 failure example.

VLM Evaluation - Failure Example #3

InternVL-2.5

Evidence #1: 08:48:40-08:48:42
The barrier arm is seen descending towards the person on the scooter.

"08:48:40": "[x_min_1, y_min_1, x_max_1, y_max_1]"
"08:48:41": "[x_min_2, y_min_2, x_max_2, y_max_2]"
"08:48:42": "[x_min_3, y_min_3, x_max_3, y_max_3]"

Answer: "D": "The person would have been hit by the descending red and white barrier arm."

GT Answer: "F": "They would have walked into the path of the moving car."

Analysis: The VLM can generate reasonable (but wrong) evidence rationale, but fail to follow the complex instruction in the prompt to generate valid temporal grounding and spatial grounding.

Figure 18. **Case Studies and Failure Analysis.** #3 failure example.