This CVPR Findings paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Distilling Counterfactual Reasoning from Language to Vision: Causal Graph Guided Post-Training for Video Understanding

Yuefei Chen[1], Jiang Liu[2], Xiaodong Lin[1], Ruixiang Tang[1†]

[1]Rutgers University, [2]Advanced Micro Devices

chen.yuefei@rutgers.edu, ruixiang.tang@rutgers.edu

Abstract

Vision Language Models (VLMs) have recently shown significant advancements in video understanding, especially in feature alignment, event reasoning, and instruction-following tasks. However, their capability for counterfactual reasoning, inferring about alternative outcomes under hypothetical conditions, remains underexplored. This capability is essential for robust video understanding, as it requires identifying underlying causal structures and reasoning about unobserved possibilities, rather than merely recognizing observed patterns. To systematically evaluate this capability, we introduce CounterVQA, a video-based benchmark featuring three progressive difficulty levels that assess different aspects of counterfactual reasoning. Through comprehensive evaluation of both state-of-the-art open-source and closed-source models, we uncover a substantial performance gap: while these models achieve reasonable accuracy on simple counterfactual questions, performance degrades significantly on complex multi-hop causal chains. To address these limitations, we develop a post-training method, ***CFGPT****, that enhances a model's visual counterfactual reasoning ability by distilling its counterfactual reasoning capability from the language modality, yielding consistent improvements across all CounterVQA difficulty levels. The project page is available at: https://github.com/AzhenHETU/CounterVQA.*

1. Introduction

Visual language models (VLMs) have recently shown remarkable progress in the field of video understanding, especially in tasks such as temporal feature alignment, complex event reasoning, and instruction following [3, 29, 35, 39, 43]. These achievements create an impression that VLMs are rapidly approaching deep cognition and high-level reasoning capabilities in dynamic worlds. However, a deeper sign of intelligence is not only recognizing and describing events that have occurred, but understanding why events occured, then reasoning alternative outcomes under different interventions. This ability, that is also called counterfactual reasoning, is to envision alternate outcomes based on different choices or events, represents the pinnacle of Pearl's Causal Hierarchy [33] and underpins core human cognitive capacities including decision-making, moral reasoning, and problem-solving. Reaching counterfactual reasoning ability is widely considered a prerequisite for building robust, generalizable AI and moving towards Artificial General Intelligence [23, 30].

Despite recent progress, several studies suggest that current VLMs may not truly reason about causality. Work on temporal reasoning and hypothetical queries has shown that VLMs often rely on superficial correlations or linguistic shortcuts rather than understanding underlying causal relations [5, 26, 45]. These findings imply that VLMs may also struggle with counterfactual reasoning, however, no existing research has systematically evaluated this ability in video settings or investigated how to improve it. Motivated by this gap, we aim to address two key questions:

How well do VLMs handle counterfactual video reasoning? Recent benchmarks have advanced the evaluation of causal reasoning in VLMs along several dimensions. Some works [7, 22, 36] evaluate causal structure inference in static images. Other works [20, 34, 44] assess temporal and compositional reasoning through language. The other works [14, 26, 27] probe physical reasoning and anomaly detection in videos.

Our work complements these efforts by introducing a benchmark CounterVQA, the first video benchmark that systematically generates and verifies counterfactual questions grounded in explicit causal graphs. Unlike previous datasets, CounterVQA employs a multi-agent generation framework that simulates human-like causal reasoning. This graph-based and agent-driven design allows precise control over causal chain complexity, enabling the con-

† Corresponding author.

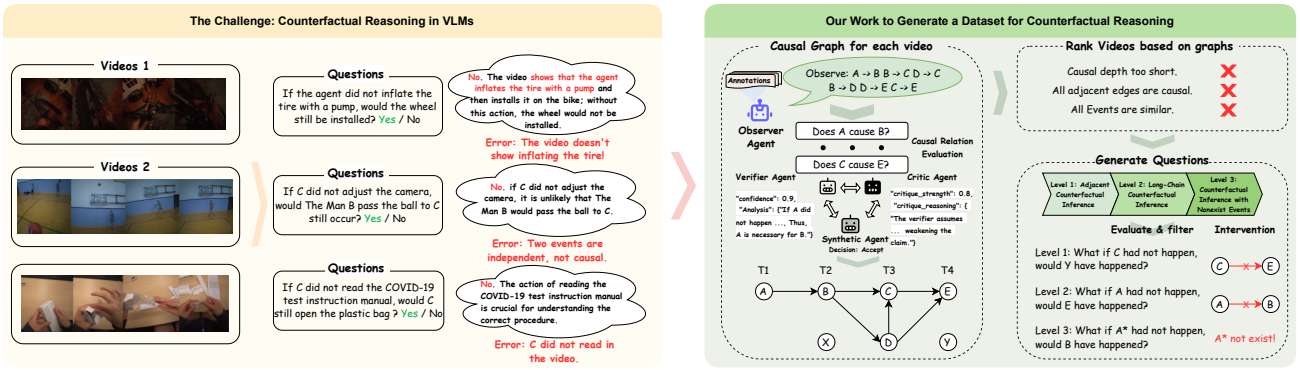


Figure 1. The left panel illustrates typical wrong cases where current VLMs misinterpret causal relations when answering counterfactual video questions. The right panel presents our dataset generation pipeline: a multi-agent system infers pairwise causal relations to build causal graphs for each video, ranks videos by causal graph complexity, and generates three levels of counterfactual questions, adjacent, long-chain, and non-existent event inference, for systematic evaluation.

struction of three progressive difficulty levels questions that rigorously test reasoning from single-step to complex multi-hop causal chains.

How to enhance VLM's counterfactual video reasoning abilities? Simply scaling up supervision or performing conventional fine-tuning is insufficient, as it tends to reinforce overfitting to observed patterns rather than fostering true causal understanding. What is needed is a causally informed training paradigm that can teach models to align video perception with intervention-based reasoning.

To address this challenge, we propose a novel post-training framework that integrates supervised fine-tuning (SFT) with reinforcement learning (RL) training [6, 13, 19] guided by causal graph rewards. The central innovation lies in introducing a Causal Graph Reward, a structured signal derived from the explicit causal graphs used in CounterVQA's multi-agent generation process. During training, the language modality acts as a teacher that excels at logical abstraction, while the visual modality learns as a student to perceive and internalize these abstract causal relations from video evidence. The causal graph reward quantifies consistency between predicted and ground-truth causal structures and it encourages the model to move beyond pattern recognition and to reason over the underlying causal mechanisms that govern event transitions. Empirical results demonstrate that this causally informed post-training improves performance across all difficulty levels in CounterVQA. In summary, our main contributions are as follows.

- We introduce **CounterVQA**, one of the largest benchmark for assessing counterfactual reasoning ability in VLMs for various situations, designed to evaluate short-chain causal relation, long-chain counterfactual reasoning, and hallucination in counterfactual reasoning.
- We propose a two-stage post-training methodology **CFGPT** guided by counterfactual causal graphs, which improves VLMs' ability to perform causal and counterfactual reasoning over video content.
- We benchmark CounterVQA using five state-of-the-art open- and closed-source models, and demonstrate the effectiveness of CFGPT, which achieves a 12.5% improvement over the leading 8B-scale VLM across all categories and difficulty levels.

2. CounterVQA: Benchmark for VLM video counterfactual reasoning

To evaluate counterfactual reasoning in VLMs, we introduce CounterVQA, a video-based benchmark with 712 videos and 3,987 QA pairs. The benchmark features a two-tier taxonomy: (1) three levels of counterfactual complexity, and (2) interaction type (Human-to-Human vs. Human-to-Object).

2.1. Counterfactual Challenge Levels

Counterfactual reasoning requires models to imagine alternative outcomes under hypothetical conditions that deviate from observed reality. To systematically assess this capability, we design a three-level hierarchy as shown in Figure 2. This hierarchy progressively increases reasoning complexity. Each level targets distinct aspects of causal understanding: from immediate causal effects to long-range temporal dependencies and resistance to plausible distractors. This hierarchical design allows us to precisely characterize current VLMs' performance in counterfactual reasoning.

2.1.1. Level 1: Adjacent Counterfactual Inference

Adjacent counterfactual inference evaluates a model's ability to reason about how intervening on one action affects its immediate successor in an event chain. This represents the most fundamental form of counterfactual reasoning, requir-

ing two distinct capabilities. First, models must differentiate causal relationships from temporal correlations, since adjacent events may co-occur without causal dependency. Second, given a validated causal link, models must predict how an intervention on the cause propagates to the effect. By testing adjacent inference, we assess whether VLMs possess the foundational ability to identify and reason over local causal dependencies, which is prerequisite for more complex multi-step reasoning tasks.

Formulation. Let $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ denote a sequence of actions observed in the video, where a_i causally precedes a_{i+1}. An adjacent counterfactual question intervenes on action a_i with an alternative a_i' and asks about the resulting change to the immediate next action:

$$\text{Query: } a_i \to a_i' \implies a_{i+1} \to ? \tag{1}$$

The model must predict the modified outcome a_{i+1}' under the counterfactual condition $a_i \to a_i'$. This requires both identifying the causal nature of the relationship $a_i \to a_{i+1}$ distinguishing it from mere temporal adjacency and reasoning about how the intervention propagates through this causal link while holding other factors constant.

Example. In Figure 2 coffee-making video, we assume the original action a_i is "pour milk into the cup" leading to a_{i+1} = "add hot water into the coffee mug", the counterfactual question asks: "If pouring milk had not happened ($\neg a_i$), would the agent still add hot water into the coffee mug?" The model must recognize that these are adjacent but independent events. That is pouring milk does not causally affect adding hot water.

2.1.2. Level 2: Long-Chain Counterfactual Inference

Long-chain counterfactual inference examines whether models can trace the cascading effects of an intervention across multiple steps in a causal sequence. Unlike adjacent inference, this level requires reasoning about non-adjacent actions where the intervention propagates through intermediate events before affecting the target outcome. This tests VLMs' capacity for multi-step causal reasoning and their ability to maintain counterfactual consistency over extended temporal horizons. These capabilities are essential for understanding complex real-world scenarios where actions have delayed or indirect consequences.

Formulation. Given an action sequence $\mathcal{A} = \{a_1, \ldots, a_n\}$ and a causal chain $a_i \to a_{i+1} \to \cdots \to a_{i+k}$ where $k \geq 2$, a long-chain counterfactual question intervenes on a_i and queries about the effect on a non-adjacent action a_{i+k}:

$$\text{Query: } a_i \to a_i' \implies a_{i+k} \to ? \tag{2}$$

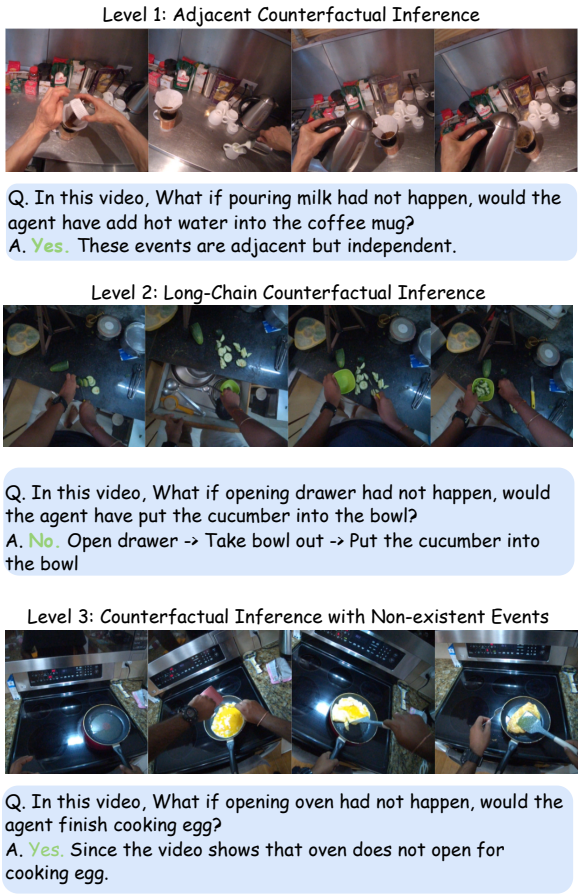


Figure 2. Representative examples from the three difficulty levels of CounterVQA. **Level 1:** Adjacent counterfactual inference requires reasoning about direct causal dependencies between consecutive events. **Level 2:** Long-chain counterfactual inference involves tracing multi-step causal relationships across several actions. **Level 3:** Counterfactual inference with non-existent events demands reasoning about hypothetical scenarios that did not occur in the observed video.

The model must reason through the intermediate causal links $a_i' \to a_{i+1}' \to \cdots \to a_{i+k}'$ to predict the final outcome. This requires tracking how changes propagate through the causal graph over multiple hops.

Example. In Figure 2 cooking preparation video, we assume a_i = "open drawer", which leads through a_{i+1} = "take bowl out" to a_{i+2} = "put the cucumber into the bowl", the counterfactual asks: "If opening drawer had not happened ($\neg a_i$), would the agent have put the cucumber into the bowl (a_{i+2})?" The model must trace the multi-step causal chain: opening drawer $\to$ taking bowl out $\to$ putting cucumber. Without the initial action, the entire chain collapses.

2.1.3. Level 3: Counterfactual Inference with Non-existent Events

This level evaluates the model's ability to reason about hypothetical events that never occurred in the observed video. A critical challenge in counterfactual reasoning is distinguishing between what actually happened and what could have happened but didn't. VLMs must not only perform counterfactual inference but also verify whether the premise of the counterfactual question is grounded in actual visual observations. This level tests whether models can correctly identify and reason about actions that are semantically plausible given the video context but factually absent from the visual evidence. By requiring models to verify event occurrence before counterfactual inference, we assess their ability to maintain factual grounding while engaging in hypothetical reasoning. It is a critical capability for reliable deployment in real-world applications where distinguishing observed facts from plausible assumptions is essential.

Formulation. Let $\mathcal{A}_{\text{true}} = \{a_1, \ldots, a_n\}$ denote the ground-truth action sequence and $\mathcal{A}_{\text{distractor}} = \{\tilde{a}_1, \ldots, \tilde{a}_m\}$ denote a set of plausible but non-existent actions that are semantically coherent with the video context. A Level 3 question presents a counterfactual scenario involving $\tilde{a}_j \in \mathcal{A}_{\text{distractor}}$:

$$\text{Query: } \tilde{a}_j \to \tilde{a}'_j \implies a_k \to ? \tag{3}$$

The correct response is to recognize that $\tilde{a}_j$ never occurred, thus rendering the counterfactual premise invalid. This tests the model's ability to distinguish $\mathcal{A}_{\text{true}}$ from $\mathcal{A}_{\text{distractor}}$ under challenging conditions where semantic plausibility conflicts with factual accuracy.

Example. In the Figure 2 cooking egg video showing only an unopened oven with an egg inside, the question asks: "If opening oven had not happened ($\neg\tilde{a}_j$, where $\tilde{a}_j$ is a non-existent event), would the agent finish cooking egg?" The model must recognize that the oven never actually opened in the video. Since the oven remained closed throughout, no cooking occurred.

2.2. Interaction Type

We divide videos into two categories based on agent interaction type, as each requires distinct causal reasoning skills. **Human-to-Human (H2H)** interactions involve collaborative activities such as basketball, soccer, or dance, where reasoning must capture social dynamics and mutual dependencies between agents. **Human-to-Object (H2O)** interactions focus on tasks where a person manipulates objects, such as cooking, climbing, or COVID-19 testing, which requires understanding of physical causality and how actions affect material states. By contrasting H2H and H2O, we assess whether VLMs comprehend both social and physical causality, revealing whether their reasoning generalizes across causal domains.

2.3. Dataset Generation

2.3.1. Dataset Collection

We build our benchmark on Ego-Exo4D [18], a large-scale multi-view video dataset comprising 5,035 videos that capture skilled human activities from synchronized egocentric and exocentric perspectives. The dataset covers diverse scenarios including physical activities such as soccer, basketball, and dance, as well as procedural tasks such as cooking and bike repair. To ensure high-quality counterfactual reasoning evaluation, we develop a three-stage pipeline to select suitable videos and generate challenging questions from random 1200 videos. This process filters the original collections down to around 700 and produces over 3,000 question-answer pairs, as detailed below.

2.3.2. Counterfactual Questions Generations

Generating high-quality counterfactual questions for videos is challenging: naive LLM-based generation often produces questions that are either too simple or lack grounding in the video's causal structure. To address this, we propose a three-stage pipeline that progressively refines question quality. First, we construct explicit causal graphs using multi-agent systems to capture the key event dependencies in each video. Second, we design graph-based complexity metrics to filter videos that are suitable for counterfactual reasoning. Finally, we employ LLMs to generate questions and use paired LLM verification to ensure answer quality. This approach increases question difficulty substantially: models' accuracy drops by 12% compared to questions generated by baseline LLM-only methods, validating the effectiveness of our pipeline.

Step1: Multi-Agent Causal Graph Construction. For each video, we employ a multi-agent system to construct a causal graph G = (V, E), where nodes V represent key events and edges E capture causal dependencies. Specifically, we use 4 specialized agents. **Observer Agent**: We first generate a set of candidate causal relations and let the LLM assign each relation a confidence score, representing the model's self-estimated likelihood that the relation is correct. We then rank all candidate relations by these confidence scores and retain only the top 20% for further evaluation by the Verifier Agent and the Critique Agent. **Verifier Agent**: Evaluate a single proposed causal link between two actions based on Commonsense Knowledge. **Critic Agent**: The goal is to find flaws in the Verifier's causal analysis. **Synthesizer Agent**: Make a final decision for this causal relational decision based on confidence scores from Verifier and Critic Agents. After that, all accepted causal relations

are aggregated into a causal graph for each video. Human evaluation indicates that roughly 90% of the extracted relations align with human judgments of causal correctness, as detailed in Appendix D.

Step2: Video Filtering via Graph Complexity Metrics. Not all videos are equally suitable for counterfactual reasoning. To ensure our benchmark contains videos with rich causal structures across all three difficulty levels, we design three graph-based metrics that directly correspond to the capabilities tested at each level. These metrics quantify different aspects of causal complexity and enable systematic filtering of suitable videos. **Adjacent Non-Causal Density (ANCD)**: To assess whether a video contains sufficient temporal correlations that must be distinguished from true causality, we compute the ratio of adjacent non-causal relationships to all adjacent relationships:

$$\text{ANCD}(G) = \frac{\#\text{ Adjacent Non-causal Relations}}{\#\text{ Adjacent Relations}} \tag{4}$$

A higher value indicates more challenging scenarios for Level 1, as models must carefully differentiate spurious temporal adjacency from genuine causal links. We retain videos with ANCD ≥ 0.2 to ensure sufficient difficulty in identifying true causal dependencies. **Causal Depth (CD)**: To evaluate a video's suitability for long-chain counterfactual reasoning, we measure the maximum path length in the causal graph:

$$\text{Causal-Depth}(G) = \max_{p \in P(G)} \text{length}(p) \tag{5}$$

where $P(G)$ denotes the set of all simple paths in the causal graph G, i.e., paths that do not revisit any node. Videos with greater causal depth contain longer causal chains, enabling multi-step reasoning questions. We select videos with Causal-Depth ≥ 3 to ensure adequate complexity for Level 2 questions. **Contextual CNDA Score**: Level 3 questions require plausible but non-occurring actions as distractors. We compute the Contextual CNDA score to quantify semantic diversity within each video: videos with diverse actions provide richer semantic contexts where additional plausible alternatives can be naturally conceived. For each action node v in the graph, we first calculate its outlier score based on semantic deviation from its task context $\tau(v)$:

$$\mathbf{c}_{\tau(v)} = \frac{1}{|C_{\tau(v)}|} \sum_{u \in C_{\tau(v)}} e(u) \tag{6}$$

where $C_{\tau(v)}$ is the set of actions sharing the same task context, and $e(u)$ is the embedding of action u. The outlier score measures how semantically atypical an action is:

$$\text{OutlierScore}(v) = 1 - \frac{e(v) \cdot \mathbf{c}_{\tau(v)}}{\|e(v)\| \cdot \|\mathbf{c}_{\tau(v)}\|} \tag{7}$$

We then average across all nodes to obtain the video-level CNDA score:

$$\text{Avg-CNDA-Score}(G) = \frac{1}{|V|} \sum_{v \in V} \text{OutlierScore}(v) \tag{8}$$

A higher CNDA score indicates that the video contains diverse actions, making it easier to construct plausible hallucinated alternatives. We retain videos with avg-CNDA-score ≥ 0.12 to ensure sufficient semantic diversity for Level 3 distractor generation. These metrics filtering reduces the candidate set from 1,200 videos to 700+, ensuring that retained videos exhibit rich causal structures necessary for challenging counterfactual reasoning evaluation.

Step3: Question Generation and Quality Control. For each filtered video, we use Deepseek-V3 to generate counterfactual questions based on the causal graph. The generation prompt instructs the LLM to identify a pivotal node in the graph, hypothesize an alternative outcome at that node, and reason about downstream effects. To ensure question quality, we implement a paired verification process. Two independent text-only LLMs (Qwen 2.5-32B) attempt to answer each generated question using only the video captions. We only keep the challenging questions which can be answered correctly by LLMs with video annotations. This verification step removes approximately 80% of generated questions, resulting in a final set of 3,000+ high-quality QA pairs. To ensure the reliability of causal graph and questions, we are conducting a validation on the dataset, including inter-annotator agreement and relation-type statistics. The detailed annotation consistency results and question-type distributions are included in the Appendix C.

3. Counterfactual Graph-based Post-Training (CFGPT)

Beyond benchmarking, we explore how to systematically enhance VLMs' counterfactual reasoning in videos. Our investigation reveals two key insights into current VLMs' limitations: **1. Cross-modal reasoning gap**: While VLMs demonstrate strong causal reasoning in text, they cannot ground these capabilities in video understanding, suggesting difficulties in transferring reasoning patterns from linguistic to visual-temporal domains [32, 41]. **2. Lack of structural guidance**: VLMs frequently produce counterfactual reasoning that either invents visual content or creates inconsistent causal chains, indicating insufficient supervision from both visual evidence and explicit causal structures. To probe the source of these limitations, we test whether reasoning improves when explicit temporal and relational cues are provided. We compare two prompting settings on Qwen-2.5-VL 32B: using only raw video versus using video with structured annotations (V+A). This allows

Teacher Setting	Accuracy (%)	Readability Score (5.0)
Video	41.9	3.2
Video+Annotations	**46.7**	**3.5**

Table 1. Comparison of outputs under video+annotation (V+A) and video-only (V) prompting. Readability metric details are in Appendix E.

us to assess whether adding causal structure yields more coherent counterfactual reasoning. As shown in Table 1, V+A consistently produces higher accuracy and more logically grounded explanations. These findings indicate that structured causal cues substantially enhance reasoning quality and motivate a training paradigm that can transfer these improvements to video-only models.

Building on this insight, we propose **Counterfactual Graph-based Post-Training (CFGPT)**, a two-stage framework for strengthening video-grounded counterfactual reasoning (Figure 1). Stage I (Cross-Modal Causal Transfer) distills textual causal reasoning into video-grounded inference using structured supervision. Stage II (Visual-Causal Reinforcement) further refines this skill with rewards from visual grounding and causal-graph alignment, ensuring both factual correctness and causal consistency.

3.1. Cross-Modal Causal Transfer Knowledge

The first stage of CFGPT enables the transfer of causal reasoning capabilities from text-based understanding to video-grounded inference. While language models excel at reasoning about causal relationships described in text, VLMs struggle to perform similar reasoning when grounded in dynamic visual content. We bridge this gap through carefully designed supervised fine-tuning that teaches the model to ground textual causal reasoning patterns in temporal video understanding. To support this process, We construct training data that explicitly connects textual causal reasoning with video content. Starting from videos with rich annotations, we leverage LLMs to generate chain-of-thought rationales that articulate causal relationships observable in the video. This produces training instances in the format of *(video, question, CoT reasoning, answer)*, where the CoT serves as a bridge connecting visual observations to causal inference, enabling cross-modal knowledge transfer. Given the base VLM denoted as f_{VLM}, we fine-tune it on the SFT data $\mathcal{D}_{SFT} = \{(v_i, q_i, c_i, a_i)\}_{i=1}^{N}$, where v_i represents the video, q_i is the question, c_i is the chain-of-thought rationale, and a_i is the answer. The training objective is:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N} [\log p(c_i|v_i, q_i) + \log p(a_i|v_i, q_i, c_i)] \quad (9)$$

This objective encourages the model to first generate coherent reasoning chains that capture the causal dependencies, and then produce accurate answers conditioned on both the video context and the reasoning process. The explicit CoT rationales serve as the crucial bridge to connect Textual and Visual Reasoning in this transfer process. By verbalizing the reasoning process grounded in visual observations, the model learns to translate its textual causal understanding into video-specific inference patterns. This enables the model to identify temporal dependencies, recognize causal events in motion, and reason about counterfactual scenarios based on what is actually shown in the video.

3.2. Visual-Causal Alignment Optimization

While the transfer learning stage establishes basic video-grounded reasoning, the model may still exhibit inconsistencies between its visual understanding and causal inference. To address this, we employ reinforcement learning with dual guidance from both visual content and explicit causal graph structures, ensuring the model's reasoning is simultaneously grounded in what it observes and structurally sound in its causal logic. For each training video, we construct a counterfactual causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that explicitly represents the causal relationships between events and entities mentioned in section 2.3 Multi-agent part. The nodes $\mathcal{V}$ correspond to key events in the video, while edges $\mathcal{E}$ encode the causal dependencies. These graphs serve as structural guidance to evaluate whether the model's reasoning aligns with correct causal pathways. We design a reward model that assesses the quality of the model's outputs along multiple dimensions:

$$R(o|v, q, \mathcal{G}) = \alpha R_{\text{causal}}(o, \mathcal{G}) + \beta R_{\text{visual}}(o, v) \quad (10)$$

where o represents the model's output (including both CoT and answer), R_{causal} evaluates alignment with the causal graph structure, and R_{visual} assesses the alignment between the VLM's understanding and the actual video content. The hyperparameters α, β balance these objectives. While the transfer learning stage establishes basic video-grounded reasoning, the model may still exhibit inconsistencies between its visual understanding and causal inference. To address this, we employ reinforcement learning with dual guidance from both visual content and explicit causal graph structures, ensuring the model's reasoning is simultaneously grounded in what it observes and structurally sound in its causal logic. Following the GRPO framework, we sample K outputs for each input and optimize through relative comparison. Through iterative sampling, dual reward evaluation, and policy updates, the model learns to consistently generate reasoning that aligns with both visual observations and causal structures. This dual supervision helps the model avoid two common failure modes: semantic related content not present in videos (R_{visual}) and making logically inconsistent causal inferences (R_{causal}), leading to more reliable video understanding.

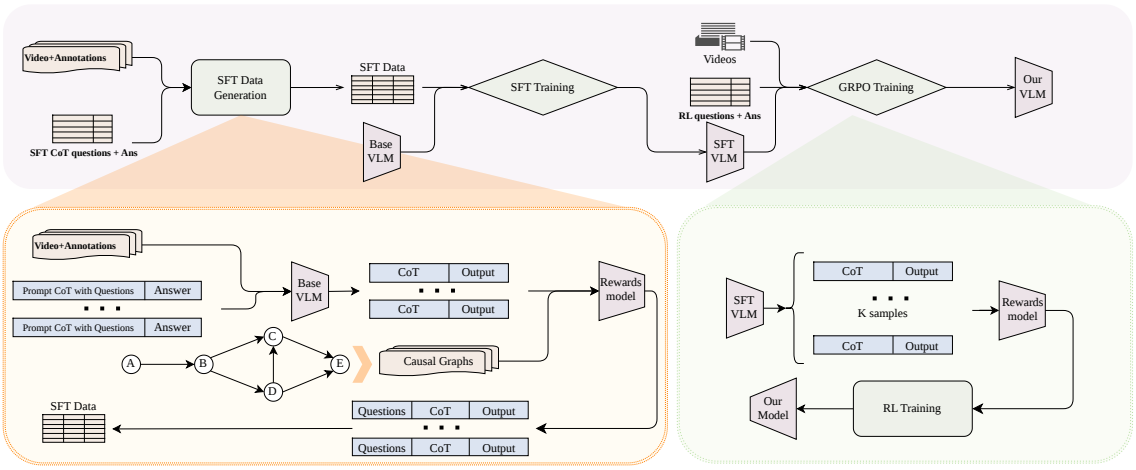


Figure 3. Overview of the CFGPT framework. **Top:** Two-stage pipeline from a base VLM to the final CFGPT model. **Left (orange):** Cross-modal causal transfer through supervised fine-tuning using V+A generated CoT data. **Right (green):** Visual-causal reinforcement via GRPO, where candidate outputs are scored by causal graph consistency and visual grounding rewards to refine counterfactual reasoning.

4. Experiment

4.1. Evaluation Settings and Models

We conduct experiments on the proposed CounterVQA benchmark to evaluate the counterfactual reasoning capability of VLMs. CounterVQA consists of three difficulty levels, Adjacent Counterfactual Inference, Long-Chain Counterfactual Inference, and Counterfactual Inference with Non-existent Events. Videos in these dataset are categorized into H2H and H2O. We evaluate different vanilla open-source models, including Qwen2.5-VL-7B, Qwen2.5-VL-32B [3], and Qwen3-VL-8B [39], and other closed source LLMs like GPT-4o [31] and Gemini-2.5 pro [17]. The details of parameters and running environments are described in Appendix A.

4.2. Main Result and Analysis

Table 2 presents the performance of all tested VLMs on CounterVQA, showing accuracy metrics as percentages across three difficulty levels and two interaction types (H2H and H2O). The results reveal several critical observations that validate both the effectiveness of our benchmark and the superiority of our proposed CFGPT method.

Our benchmark effectively exposes limitations in current VLMs. Despite recent advances, most existing VLMs demonstrate significant limitations in counterfactual reasoning. ChatGPT-4o, a leading closed-source model, achieves only 6.2% average accuracy. Even the larger Qwen-2.5-VL-32B model struggles with an average accuracy of 39.4%. These results confirm that counterfactual reasoning in videos remains a fundamental challenge that cannot be solved by simply scaling model parameters, validating the necessity and difficulty of our benchmark. Beyond overall performance, we further analyze how different types of interactions affect model reasoning capabilities. Comparing the two interaction types, we observe the results do not show a consistent advantage for either Human-to-Human or Human-to-Object reasoning across models. Some models perform better on object-centered interactions, while others handle social interactions more effectively. This variability indicates that current VLMs struggle to generalize causal reasoning uniformly across different interaction types.

CFGPT significantly improves counterfactual reasoning capabilities. Our method demonstrates substantial and consistent improvements across all difficulty levels and interaction types. Qwen-3-VL-8B + SFT alone improves the vanilla model from 60.1% to 69.3% (+9.2 points), demonstrating the effectiveness of cross-modal causal transfer learning. More importantly, adding GRPO further boosts performance to 72.6% (+12.5 points over vanilla), achieving the best results across all models. The improvement is particularly pronounced on harder levels: on Level 3 H2O tasks, our full method achieves 77.4% compared to 66.4% for the vanilla model (+11 points), demonstrating that visual-causal dual reinforcement is especially effective for complex counterfactual scenarios. However, the effectiveness of CFGPT is not uniform across all base models. Model capacity influences the effectiveness of CFGPT. CFGPT yields only modest gains on Qwen-2.5-VL-7B (8.3% $\rightarrow$ 9.3%) but substantially improves Qwen-3-VL-8B (60.1% $\rightarrow$ 72.6%), indicating that a minimum level of video understanding is required for effective counterfac-

	Human to Human				Human to Object				Total			
Models	**Level 1**	**Level 2**	**Level 3**	**Avg.**	**Level 1**	**Level 2**	**Level 3**	**Avg.**	**Level 1**	**Level 2**	**Level 3**	**Avg.**
Open-source (Vanilla)												
Qwen-2.5-VL 7B	13.3	11.5	3.1	9.1	12.1	9.9	2.5	8.1	12.3	10.3	2.7	8.3
Qwen-2.5-VL 32B	28.6	21.4	64.3	36.7	33.5	51.5	41.9	39.7	33.0	46.2	44.0	39.4
Qwen-3-VL 8B	45.1	54.5	52.6	49.1	56.7	63.7	66.4	61.9	54.6	62.1	65.3	60.1
Closed-source (API)												
ChatGPT-4o	12.5	15.8	8.3	12.7	9.0	6.5	0.7	3.9	9.6	8.0	1.3	6.2
Gemini-2.5-Pro	57.1	25.0	40.0	43.8	39.3	42.4	42.6	41.3	41.1	40.5	42.4	41.5
Open-source (Finetuned / Ours)												
Qwen-2.5-VL 7B + SFT	12.5	15.4	3.6	10.7	12.8	9.9	3.3	8.7	12.7	11.2	3.4	9.1
Qwen-3-VL 8B + SFT	76.3	54.5	**73.7**	**72.0**	65.9	71.4	70.6	68.8	68.9	67.9	71.0	69.3
Qwen-2.5-VL 7B + SFT + RL	12.1	15.4	3.5	10.3	12.8	9.9	4.2	9.0	12.7	11.2	4.1	9.3
Qwen-3-VL 8B + SFT + RL	**77.4**	**57.1**	63.2	68.8	**70.6**	**73.9**	**77.4**	**73.8**	**70.1**	**71.6**	**76.0**	**72.6**

Table 2. Performance of various VLMs on CounterVQA across three difficulty levels and two interaction types. Our finetuned models (SFT and SFT+RL) achieve the highest accuracy overall.

tual reasoning transfer. Qwen-3-VL-8B's higher baseline provides a better foundation for cross-modal transfer and reinforcement stages.Specialized post-training is more important than scale. Qwen-3-VL-8B (60.1%) already surpasses larger models such as Qwen-2.5-VL-32B (39.4%) and Gemini-2.5-Pro (41.5%), highlighting that architecture and training objectives matter more than parameter count. With CFGPT, Qwen-3-VL-8B reaches 72.6%, outperforming these larger models by over 30 points. It shows that targeted post-training is essential for strengthening counterfactual reasoning in videos, whereas simply scaling model size or relying on general purpose training is insufficient. Additional ablations study and error analysis are provided in Appendix B.

5. Related Work

Vision Language Models. Early large-scale VLMs connect pretrained vision encoders with LLMs and learn to align modalities through lightweight adapters or instruction tuning. Representative lines include Flamingo, which interleaves image/video tokens with text for few-shot multimodal prompting [2]; PaLI/PaLI-X that scale both the vision and language components jointly [8, 9]; BLIP-2 that freezes the image encoder and LLM while learning a Querying Transformer as a bridge [25]; and instruction-tuned systems such as LLaVA that distill multimodal instruction-following behaviors [28]. Recent open models introduce dynamic-resolution tokenization and unified image–video processing, improving long-video comprehension [3, 35].

Counterfactual Reasoning in VLMs. Counterfactual reasoning probes whether a model can answer "what if" questions beyond mere pattern association. Recent datasets for images (e.g., C-VQA) explicitly inject counterfactual presuppositions and show that state-of-the-art VLMs still struggle on counterfactual queries despite strong perceptual skills [46]. Related benchmarks [1, 42] demonstrate that models often fail when causal, hypothetical, or distribution-shift reasoning is required. Contemporary work also explores counterfactual fine-tuning objectives for VLMs to improve semantic discrimination and causal robustness, indicating that exposing models to minimally edited pairs can help disentangle spurious cues [45]. Overall, the gap between visual recognition and robust counterfactual inference remains a key challenge.

Benchmarks for Video Understanding and Counterfactual Reasoning. A line of video reasoning benchmarks stresses temporal/causal structure. CLEVRER introduces synthetic collision videos with descriptive, explanatory, predictive, and counterfactual questions, catalyzing research on causal video QA [40]. CATER focuses on compositional actions and temporal reasoning [16]. NExT-QA moves from description to explanation, grouping questions into causal and temporal categories over real videos [37]. For physically grounded causal reasoning, the Physion benchmark evaluates predictions of physical outcomes from videos [4]. Very recent video QA benchmarks foreground hypothetical/anticipation/counterfactual queries on real-world scenarios, and consistently report a large human–model gap [11, 15]. Causal reasoning benchmarks like CLADDER [21] and CounterBench [10] show LLMs' weaknesses in multi-step and counterfactual inference. Recent benchmarks such as Causal-VidQA, Black Swan, and FunQA [12, 24, 38] extend video understanding beyond recognition. In contrast, CounterVQA focuses specifically on video-grounded counterfactual reasoning with explicit causal graphs and progressive difficulty control.

6. Conclusion

We introduce CounterVQA, a benchmark for evaluating counterfactual reasoning in VLMs. It reveals substantial gaps in current models and motivates CFGPT, a post-training framework that aligns reasoning with video-grounded causal structures. Extensive experiments demonstrate consistent improvements across models.

Acknowledgements

We acknowledge the computational resources provided through NSF ACCESS and the support of Advanced Micro Devices, Inc. through the AMD University Program's AI & HPC Cluster.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 8

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 7, 8

[4] Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021. 8

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 1

[6] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025. 2

[7] Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. Cello: Causal evaluation of large vision-language models. *arXiv preprint arXiv:2406.19131*, 2024. 1

[8] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 8

[9] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 8

[10] Yuefei Chen, Vivek K Singh, Jing Ma, and Ruixiang Tang. Counterbench: A benchmark for counterfactuals reasoning in large language models. *arXiv preprint arXiv:2502.11008*, 2025. 8

[11] Zhenfang Chen, Shilong Dong, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Compositional physical reasoning of objects and events from videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 8

[12] Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24201–24210, 2025. 8

[13] Alan Dao and Dinh Bach Vu. Alphamaze: Enhancing large language models' spatial intelligence via grpo. *arXiv preprint arXiv:2502.14669*, 2025. 2

[14] Hang Du, Sicheng Zhang, Binzhu Xie, Guoshun Nan, Jiayang Zhang, Junrui Xu, Hangyu Liu, Sicong Leng, Jiangming Liu, Hehe Fan, et al. Uncovering what why and how: A comprehensive benchmark for causation understanding of video anomaly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18793–18803, 2024. 1

[15] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 8

[16] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. *arXiv preprint arXiv:1910.04744*, 2019. 8

[17] Google DeepMind. Gemini pro. https://deepmind.google/models/gemini/pro/, 2024. Accessed: 2025-11-11. 7

[18] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 4

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[20] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 1

[21] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065, 2023. 8

[22] Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. Causalvlbench: Benchmarking visual causal reasoning in large vision-language models. *arXiv preprint arXiv:2506.11034*, 2025. 1

[23] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. 1

[24] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 8

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 8

[26] Zongxia Li, Xiyang Wu, Guangyao Shi, Yubin Qin, Hongyang Du, Tianyi Zhou, Dinesh Manocha, and Jordan Lee Boyd-Graber. Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding. *arXiv preprint arXiv:2505.01481*, 2025. 1

[27] Disheng Liu, Yiran Qiao, Wuche Liu, Yiren Lu, Yunlai Zhou, Tuo Liang, Yu Yin, and Jing Ma. Causal3d: A comprehensive benchmark for causal learning from visual data. *arXiv preprint arXiv:2503.04852*, 2025. 1

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 8

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 1

[30] Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020. 1

[31] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2025-11-11. 7

[32] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, 2022. 5

[33] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 1

[34] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 8

[36] Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. Timecausality: Evaluating the causal ability in time dimension for vision language models. *arXiv preprint arXiv:2505.15435*, 2025. 1

[37] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 8, 4

[38] Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024. 8

[39] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 7

[40] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 8

[41] Weihao Yu, Zhengyuan Yang, Lingfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024. 5

[42] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 8

[43] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1

[44] Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024. 1

[45] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025. 1, 8

[46] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21853–21862, 2024. 8