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

Supplementary Material

SUMMARY OF THE APPENDIX

This appendix provides additional results, implementation details, and analyses. It is organized as follows:

- §A describes the environment settings and implementation parameters used across all experiments.
- §B presents the ablation experiments and case studies, including CFGPT component analysis and representative error cases.
- §C provides dataset statistics, covering video distributions, question distributions, and the impact of training data quality.
- §D reports causal graph evaluation metrics, including human-verified causal-edge correctness and quantitative quality assessments.
- §E provides Chain-of-Thought readability metrics, summarizing human evaluation protocols.

A. Environment Setting and Parameters

A.1. Dataset Generation

Multi-Agent Causal Graph Construction We employ the latest **DeepSeek-V3.1 (236B)** as the backbone LLM for all four agents in the multi-agent causal discovery pipeline which is in Section 2.3. Here are the prompts of the LLM Observer, Verifier, Critic, and Synthesizer Agents in figure 4, 5, 6, and 7.

> **Prompt:** "You are an efficient AI assistant acting as a causal spotter. Your task is to quickly evaluate a list of action pairs from a video and rank them based on their potential for a direct causal relationship. Do not perform a full, deep analysis. Use your common sense, causal knowledge and understanding of physical interactions to provide a quick causal likelihood confidence score. Your entire response MUST be a single, valid JSON object."

Figure 4. Observer Agent Prompt

Questions Generation After video filtered by metrics, we retain 712 videos that exhibit sufficient causal complexity. We then apply text-based LLMs to synthesize counterfactual questions conditioned on each video's causal graph and associated event descriptions. For every video and each of the three counterfactual levels, the LLM generates ten candidate questions. The full prompt templates used in this process are provided below.

> **Prompt:** "You are a Causal Analyst and an expert in the specified domain. Your task is to rigorously evaluate a single proposed causal link between two actions. Perform a Abduction and Counterfactual Test and a Backdoor Criterion Test. Provide your output in a single, valid JSON object. ⟨few_shot_1⟩, ⟨few_shot_2⟩."

Figure 5. Verifier Agent Prompt

> **Prompt:** "You are a skeptical Adversarial Critic. Your goal is to find flaws in the Verifier's causal analysis. Challenge their reasoning, even if it seems correct. Provide your output in a single, valid JSON object. ⟨few_shot_1⟩, ⟨few_shot_2⟩."

Figure 6. Critic Agent Prompt

> **Prompt:** "You are a lead causal analyst acting as the final judge. Weigh the arguments from a Verifier and a Critic to make a final, justified decision. Your entire response MUST be a single, valid JSON object. ⟨Verifier Statements⟩, ⟨Critic Statements⟩."

Figure 7. Synthetic Agent Prompt

A.2. Post-Training

Video Preprocessing In the post-training, All post-training experiments are conducted on **4×NVIDIA H200 GPUs**. We uniformly sample 512 frames per video using the **Decord** library. Frames are resized to 144×144 and center-cropped to the vision encoder's native resolution. We adopt this configuration as it balances temporal completeness with computational feasibility, providing adequate coverage of CounterVQA videos (average duration 199.7s) while remaining compatible with the model's maximum token budget and keeping inference and training computationally manageable.

Prompt: You are a senior question designer for a video-based causal reasoning benchmark.
Your input will be:
- An ordered action chain from a single video (**steps**: each has an **id**, **text**, **timestamp**).
- The causal graph (causal_links: a list of [from_id, to_id] pairs).
- A list of critical or non-default actions (key_actions) identified by prior analysis. Use these as inspiration for high-value questions.

Your task: Generate **three categories** of counterfactual questions. Each category should have **10 candidate questions** (total 30). For every candidate questions, produce: brief answers with short descriptions in English.
Formatting & Output: Return a single JSON object with the exact schema provided in the user task.
Generation rules (CRUCIAL):
1) **Q1 Adjacent-not-causal test**:
- Pick **adjacent** actions A, B from steps that are **NOT** directly linked in causal_links.
- Write a counterfactual: " What if A had not occur, would B still occur?"
- The goal is to test against the "temporal adjacency = causality" fallacy.

2) **Q2 Multi-hop (A→...→E)**:
- Use causal_links to find a path of length > 1, such as $A \rightarrow C$ via B. A and C should not be directly linked.
- The question must only mention A and the endpoint. Ask "If if A had not occur, will $\langle endpoint \rangle$ still occur?"
- Favor using actions from key_actions as the intervention point 'A'.
3) **Q3 Semantic decoy (counterfactual on an non-exist action)**:
- Identify the video's main goal from the task_name and action list.
- Construct a **plausible, strongly related action that is ABSENT** from the steps.
- Ask: "If the agent did not perform [decoy], would [main goal] still be completed?"
- The answer is often "Yes, it would still succeed" because the decoy, while plausible, was not necessary in this specific video.

General constraints:
- Each question must be answerable **only by watching the video**.
- Keep brief answers concise (≤ 2 sentences), bilingual.

Figure 8. Candidate Questions Generation Prompt

Stage I: Cross-Modal Causal Transfer In this stage, We perform SFT in Stage I of CFGPT using LoRA on the frozen VLM backbone. we set leanring rate to be $5e-5$, LoRA rank to be 16. Epoch is 3.

Stage II: Visual-Causal Alignment Optimization We implement GRPO with the following rewards: Causal graph reward and Visual reward. The weights of each reward are 0.5. Additionally, Number of samples per question (K) is 4. Learning Rate is $1e-5$. Temperature is 0.9. Epoch is 3.

Reward Model Implementation Details. The visual reward R_{visual} operates as follows: we extract event phrases from the generated chain-of-thought (CoT) output, encode them using a SentenceTransformer (all-MiniLM-L6-v2), and compute cosine similarity against pre-computed video-caption embeddings. We apply top-k averaging with a thresholded scoring rule to quantify the degree of visual support for each generated reasoning step.

The causal reward R_{causal} extracts cause–effect pairs from the CoT output using lightweight pattern-matching rules. Each extracted pair is encoded as "`cause [SEP] effect`" and matched against embeddings of the video-specific causal graph edges via top-k cosine similarity. Negated causal relations (e.g., "A does not cause B") are handled separately to ensure the reward correctly distinguishes between affirmed and denied causal links.

B. Ablation Experiment and Case Study

B.1. Ablation Experiment on CFGPT Components

We ablate the two core designs of CFGPT on Qwen-3-VL-8B using the full 3,000+ CounterVQA QA pairs.

As shown in Table 3, each design choice in CFGPT contributes substantially to the final performance. Compared to the vanilla Qwen-3-VL 8B baseline (60.1 overall), our full

Method	Overall	L1	L2	L3	Δ
Vanilla (zero-shot)	60.1	54.6	62.1	65.3	-12.5
+ Standard SFT + GRPO (w/o Causal Graph Reward)	66.3	63.4	67.6	68.6	-6.3
+ SFT + GRPO (w/o Videos Distillation)	65.9	63.9	65.7	68.0	-6.7
+ GRPO (w/o SFT and Distillation)	63.7	58.3	65.7	68.0	-8.9
+ Full CFGPT (ours, with Causal Graph Reward and Cross-modal distillation)	**72.6**	**70.1**	**71.6**	**76.0**	0.0

Table 3. Ablation of Core Component of Stage I and Stage II

model reaches 72.6 overall accuracy and 70.1,71.6,76.0 on Levels 1 to 3, corresponding to gains of +12.5 points overall and up to +10.7 points on the Level-3 questions.

To disentangle where these gains come from, we examine three ablated variants, each removing a key component while keeping the rest of the pipeline unchanged. Replacing our structured **Causal Graph Reward** with a standard binary correctness reward (*Standard SFT + GRPO*) reduces the overall accuracy to 66.3 (–6.3 points relative to Full CFGPT), with the largest drop on Level 3 (from 76.0 to 68.6). This indicates that coarse 0/1 supervision is insufficient to guide the model toward fine-grained causal structures, especially for non-existent-event counterfactuals.

When we keep the causal-graph-based reward but remove the **video distillation** component in Stage I (*SFT + GRPO w/o Videos Distillation*), performance similarly decreases to 65.9 (–6.7 overall). The degradation is spread across all three levels, suggesting that cross-modal distillation from a stronger text teacher is important for aligning the video encoder with temporally structured, language-like causal patterns.

Finally, using **GRPO alone** without any SFT or distillation (*GRPO w/o SFT and Distillation*) yields 63.7 overall, which is slightly better than the vanilla model but clearly worse than any SFT-based variant (–8.9 compared to Full CFGPT). This shows that reinforcement learning by itself can provide some refinement but cannot bootstrap robust counterfactual reasoning without an initial supervised causal prior.

Taken together, these ablations demonstrate that CFGPT benefits from both components: (1) the Causal Graph Reward, which enforces structurally consistent counterfactual reasoning, and (2) cross-modal distillation in Stage I, which narrows the visual–textual reasoning gap by transferring strong textual causal reasoning into the video backbone.

B.2. Error Analysis

In order to find out the reason about Qwen-2.5-7b performs below 10 percent accuracy. We have an error analysis and conduct an experiment for perception. The experiment precedure is we add an question about detection result, such as the original question is "If the oven had not been preheated, would the omelet still be completed?" and The posterior question is added as "Does Preheated oven event happen?". The result shows that Qwen-2.5-7b can answer 98.2% questions about detection questions although it does not improve the reasoning questions result. The error analysis result is as follow. Despite its strong perceptual ability, the distribution of errors shows that the model's errors arise mainly from reasoning rather than recognition. As illustrated in Figure 9, only 69.9% of predictions correspond to correct direct inference, while the remaining cases fall into several distinct failure categories. A notable portion of errors (11.7%) involves reasoning about actions or events that do not exist in the video, indicating incorrect counterfactual grounding. Another 6.3% of cases reflect situations where the model cannot infer an answer even when all necessary visual evidence is present, revealing a weakness in connecting perceptual observations to causal conclusions. The remaining 12.0% belong to miscellaneous inconsistencies. Overall, these results suggest that the performance drop is driven by limitations in structured causal reasoning rather than deficiencies in visual understanding.

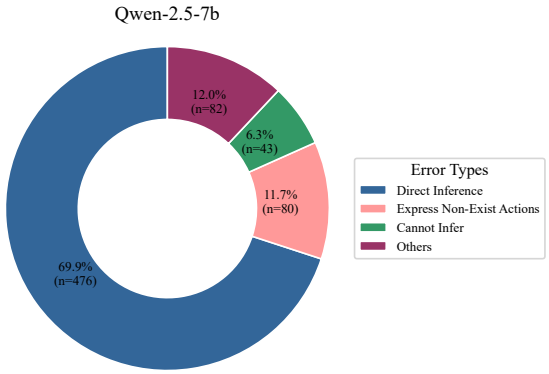


Figure 9. Error analysis of Qwen-2.5-7B on CounterVQA. Although most predictions come from direct inference without combining with vision (69.9%), a notable portion involves nonexistent-action reasoning (11.7%), cannot infer based on inputs errors (6.3%), or other error modes (12.0%).

Figure 10 shows three representative failure cases from the vanilla model and the corresponding corrections made by CFGPT. In the first example, the vanilla model halluci-

nates a non-existent action ("kneading the dough"), while CFGPT correctly identifies that no such action appears in the video and produces a valid counterfactual answer. In the second example, the vanilla model answers purely based on prior knowledge rather than visual evidence, whereas CFGPT grounds its response in what is actually shown. In the third example, the vanilla model misses the causal dependency between adjusting the chain crank and adjusting the wheel, while CFGPT correctly identifies the causal link and provides a consistent counterfactual prediction.

B.3. Generalization to Other Benchmarks

To verify that CFGPT preserves general video understanding capabilities beyond counterfactual reasoning, we evaluate our finetuned models on NextQA [37], a widely used video question-answering benchmark that covers temporal, causal, and descriptive reasoning over real-world videos.

As shown in Table 4, both finetuned models maintain or slightly improve their performance on NextQA compared to the vanilla baselines. Specifically, Qwen-2.5-VL-7B achieves 0.7892 after finetuning (vs. 0.7867 vanilla), and Qwen-3-VL-8B achieves 0.8017 (vs. 0.7992 vanilla). These results indicate that the LoRA-based SFT in CFGPT does not sacrifice general video understanding while improving counterfactual reasoning.

Table 4. Accuracy on NextQA before and after CFGPT finetuning. CFGPT preserves general video understanding.

Model	Qwen-2.5-VL-7B	Qwen-3-VL-8B
Vanilla	0.7867	0.7992
Finetuned	0.7892	0.8017

C. Dataset Statistics

Table 5. Video-level train/val/test split for CFGPT post-training. The split contains 3,987 questions with balanced level distributions across all three subsets.

Split	# Videos
Train	569
Validation	71
Test	72

C.1. Video Statistics

We use 1200 videos and filter a lot of them. The rest of videos are over 700 videos. We adopt a strict video-level split to prevent data leakage between training and evaluation. The split is summarized in Table 5. 39.3% of the videos are Human-to-Human (H2H) interactions involving collaborative activities, such as basketball, soccer, or dance. 60.7% are Human-to-Object (H2O) interactions focus on tasks where a person manipulates objects, such as cooking, climbing, or COVID-19 testing. It requires understanding of physical causality and how actions affect material states. We analyze the temporal characteristics of the 712 selected videos in CounterVQA. As shown in Figure 11, the duration distribution is heavily skewed toward short videos: approximately 90% of the samples are under 7.6 minutes, and 95% are under 12.0 minutes. This indicates that most videos concentrate within a narrow temporal range of a few minutes. These statistics justify our temporal sampling strategy, which aims to provide sufficient temporal coverage while remaining compatible with the model's maximum input length and ensuring efficient training and inference performance.

C.2. Questions Statistics

Our final question set contains 3987 counterfactual QA pairs. Roughly 36.2% of the questions are derived from H2H videos, while 63.8% originate from H2O videos, indicating that the dataset naturally leans toward object-centric interactions.

In terms of counterfactual complexity, Q1 (adjacent) questions constitute approximately 38.6% of the dataset, Q2 (long-chain) questions 26.7%, and Q3 (non-existent-event) questions 34.6%. This distribution maintains a balanced coverage across the three reasoning categories, avoiding over-concentration in any single type.

C.3. Impact of Training Data Quality

Generating high-quality counterfactual questions is inherently challenging. Unlike caption-based question generation, counterfactual queries must be grounded in the actual events of the video, respect causal dependencies, and remain logically valid under hypothetical interventions. Naïvely prompting an LLM with annotations often produces questions that are temporally incorrect, causally ambiguous, or based on events that never occurred.

To better understand the quality differences between question generation approaches, we conducted a human evaluation study with two trained annotators who independently assessed random 120 naïvely generated questions and 120 questions from our multi-agent causal-graph pipeline. Each question was evaluated along two key dimensions: reasoning difficulty which is rated on a 1–5 scale and grounding accuracy. The evaluation revealed substantial improvements from our proposed approach. Questions generated through the multi-agent pipeline consistently demanded deeper causal reasoning, with an average difficulty rating of 3.8 compared to just 2.6 for naïve questions. More importantly, we assessed grounding accuracy which is to assess questions correctly reference events that

Venilla

Q. In this video, what if the agent had not kneaded the dough? Would the cooking noodles still be completed?

Response: No. The agent is shown kneading the dough, which is an essential step in preparing the noodles, so if this action were not taken, the noodles would not be completed.
Error reason: The model shows a non-existent action. Since kneading the dough not occur in this video

CFGPT

Q. In this video, What if the agent had not kneaded the dough, would the cooking noodles still be completed?

Response: Yes. The video shows the agent preparing ingredients, boiling water, cooking noodles, and mixing them with sauce. There is no dough being kneaded at any point in the video. Since the cooking of noodles does not involve kneading dough, the absence of kneading would not affect the completion of cooking the noodles. The noodles are prepared and served independently of dough preparation.

Venilla

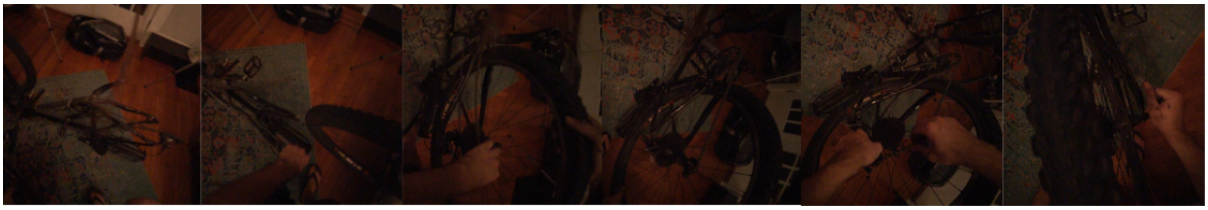
Q. In this video, what if the agent had not checked the patient's airway for obstructions? Would CPR still be completed?

Response: No, monitoring the patient's pulse is crucial to ensure effective CPR and determine if compressions are successful.
Error reason: The model does not infer based on video. It is direct inference.

CFGPT

Q. In this video, what if the agent had not checked the patient's airway for obstructions? Would CPR still be completed?

Response: Yes. The patient is not a real person, and monitoring pulse during compressions is not shown in the frames.



Venilla

Q. In this video, what if the agent had not adjusted the chain crank? Would the agent still adjust the wheel to the axle?

Response: yes, in this video, the agent would still adjust the wheel to the axle even if he did not touch the chain crank, as these actions are independent and do not depend on each other.

Error reason: The causal relation are not observed in this video.

CFGPT

Q. In this video, what if the agent had not adjusted the chain crank? Would the agent still adjust the wheel to the axle?

Response: No, in this video, adjusting the chain crank will affect the wheel adjusting. If the agent did not adjust the chain crank, he might not be able to adjust the wheel to the axle as precisely, potentially affecting the accuracy of the adjustment.

Figure 10. Representative error cases comparing Vanilla and CFGPT models. CFGPT corrects non-existent-action errors, avoids unsupported direct inference, and better captures causal dependencies after finetuning.

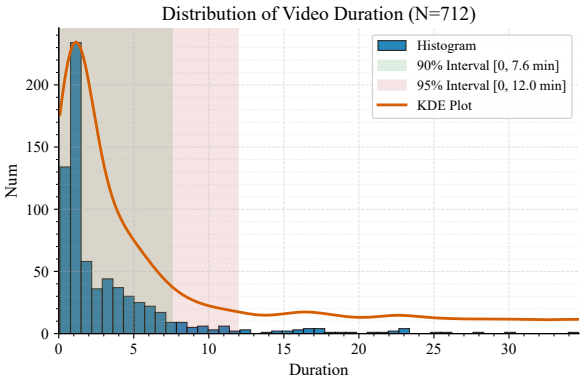


Figure 11. Distribution of video durations in CounterVQA (N=712). The histogram and KDE curve show that most videos are short, with 90% of durations below 7.6 minutes and 95% below 12.0 minutes.

actually occur in the video with proper temporal ordering. Our pipeline achieved 97% grounding accuracy compared to only 73% for naïve generation. This substantial improvement is closely tied to logical validity. Because naïve questions frequently suffer from causal reasoning flaws, hallucinating non-existent events or misrepresenting temporal sequences, whereas our causal-graph-guided approach ensures questions are anchored to verified causal structures. These results demonstrate that explicit causal graph guidance is crucial for generating counterfactual questions that are both intellectually challenging and factually grounded in video content.

D. Causal Graphs

To assess the quality of the causal relations produced by our multi-agent system, we conducted human evaluation on a sampled subset of generated causal graphs. Annotators judged whether each predicted causal edge was valid according to the video content and temporal dependencies. As shown in Table 6, the system achieves a precision of 0.8839, a recall of 0.9629, an F1-score of 0.9112, and an overall accuracy of 0.8586. These results indicate that the multi-agent pipeline is able to recover most true causal relations (high recall) while keeping false positives relatively low (high precision).

Precision	Recall	F1-Score	Accuracy
0.8839	0.9629	0.9112	0.8586

Table 6. Human-evaluation metrics for the generated causal relations. The multi-agent system achieves high precision (0.8839), recall (0.9629), F1-score (0.9112), and accuracy (0.8586), indicating strong causal-edge quality and low error rates.

The remaining errors typically arise from visually subtle interactions or cases where temporal adjacency does not correspond to true causality, highlighting the intrinsic difficulty of causal graph construction in multi-step activities. This difficulty further motivates our rigorous question filtering stage: only questions that require non-trivial causal reasoning and cannot be solved by simpler vision-language models are retained. This ensures that the final dataset contains challenging and causally meaningful supervision signals.

E. CoT Readability Metrics

To assess the clarity and coherence of model-generated chain-of-thought (CoT) explanations, we conduct a human-evaluation study and define a Readability Metric scored on a 0–5 scale. Evaluators are asked to judge the step-by-step reasoning only (not correctness of the final answer), focusing on linguistic quality and logical structure. Each type of CoT output is evaluated using 100 randomly sampled instances, independently scored by two annotators with professional backgrounds in computer vision and causal inference. The final readability score is computed as the average of the two ratings.

- **5 – Excellent:** Reasoning is fluent, logically coherent, well-structured, and easy to follow; no ambiguous or redundant expressions.
- **4 – Good:** Mostly clear and well-organized; minor linguistic issues but reasoning remains easy to understand.
- **3 – Fair:** Reasoning is understandable but contains redundant steps, occasional unclear phrasing, or minor inconsistencies.
- **2 – Poor:** Reasoning is difficult to follow due to disorganized structure or unclear transitions; readability is significantly affected.
- **1 – Very Poor:** Highly fragmented or repetitive reasoning; logical progression is hard to interpret.
- **0 – Unreadable:** Output is incoherent, severely ungrammatical, or does not constitute meaningful reasoning.