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.

Causal Chain-Guided Reasoning for Modular and Explainable Causal-Why Video Question Answering

Paritosh Parmar[1,2*] Eric Peh[1,2*] Basura Fernando[1,2,3]

[1]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

[2]Centre for Frontier AI Research, Agency for Science, Technology and Research, Singapore

[3]College of Computing and Data Science, Nanyang Technological University, Singapore

Abstract

Existing Causal-Why Video Question Answering (VideoQA) models often struggle with higher-order reasoning, relying on opaque, monolithic pipelines that entangle video understanding, causal inference, and answer generation. These black-box approaches offer limited interpretability and tend to depend on shallow heuristics. We propose a novel, modular paradigm that explicitly decouples causal reasoning from answer generation, introducing natural language causal chains as interpretable intermediate representations. Inspired by human cognitive models, these structured cause-effect sequences bridge low-level video content with high-level causal reasoning, enabling transparent and logically coherent inference. Our two-stage architecture comprises a Causal Chain Extractor (CCE) that generates causal chains from video-question pairs, and a Causal Chain-Driven Answerer (CCDA) that derives answers grounded in these chains. To address the lack of annotated reasoning traces, we introduce a scalable method for generating accurate causal chains from existing datasets. We construct human verified causal chains for 46K samples. We also propose CAUCO*, a new evaluation metric for causality-oriented captioning. Experiments on three large-scale benchmarks demonstrate that our approach not only outperforms state-of-the-art models, but also yields substantial gains in explainability, user trust, and generalization—positioning the CCE as a reusable causal reasoning engine across diverse domains. Dataset and Code:* `https://github.com/LUNAProject22/ChainReaction`.

1. Introduction

Understanding the motivations behind human actions is crucial for developing advanced systems for nuanced behavior analysis. Human actions are shaped by factors such as personal experience, emotion, social context, and culture. This complexity requires uncovering underlying causes. In this context, Causal Video Question Answering (Causal-Why VideoQA) asks models not only to recognize events but to explain why they occur—demanding higher-order reasoning beyond descriptive QA [13, 23, 38, 54].

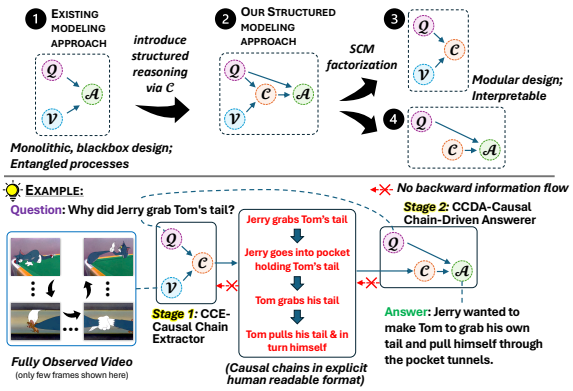

Figure 1. **(Top) Concept.** **(1)** Existing Video ($\mathcal{V}$) Question ($\mathcal{Q}$) Answer ($\mathcal{A}$) approaches through the lens of structural causal models (SCMs), highlighting their monolithic and black-box nature. **(2)** In contrast, we propose a principled departure from this paradigm: leveraging the Causal Reasoning Trace ($\mathcal{C}$), a structured intermediate representation based on natural language causal chains. We factorize this SCM into two SCMs **(3,4)**—enabling structured video understanding, reasoning, and inference—leading to superior explainability and performance. **(Bottom) Example.** *Please zoom in for the best view.*

Existing Causal-Why VideoQA models often reason from incomplete evidence or rely on shallow heuristics (*e.g.*, matching action verbs or object nouns in vision-language embedding spaces [38, 43, 52]). These models entangle video understanding, reasoning, and answer generation into one monolithic process, making their reasoning opaque and error-prone. Many high-performing vision language models (VLMs) also operate as black boxes, offering limited interpretability into their decisions.

*Equal contribution.

In this work, we make the case that reasoning and answering should be explicitly decoupled and modularized. We introduce a new paradigm in which causal reasoning & answer generation are handled by separate modules that communicate via natural language causal chains [18, 41]—structured sequences of cause-effect events that serve as intermediate reasoning steps (Figure 1). Causal chains or causal reasoning traces provide a logically coherent bridge between low-level video content and high-level causal understanding [48]. We formulate them as *linear* natural-language sequences to capture the observable steps linking a cause (*e.g.*, character's intention) to its effect, ensuring that answers are grounded in the video's causal progression. To our knowledge, ours is the first approach to explicitly use causal chains as interpretable intermediates, inspired by their role in human cognition and scientific explanation [16].

At the core of our method lies the integration of Structural Causal Models (SCMs) and Chain-of-Thought (CoT) reasoning, which enables a principled, structured decomposition of the VideoQA task. This synergy allows us to model causality in a robust and interpretable way, going beyond latent embeddings to explicitly capture causal semantics.

Our model consists of two stages: **1)** a causal chain extractor—CCE—(Figure 1(3)), and **2)** a causal chain-driven answerer—CCDA—(Figure 1(4)). The CCE model learns to extract causal chains from video, conditioned on the causal-why questions. The CCDA model learns to generate answers to causal-why questions based on the extracted causal chains. Distinct from general causal inference models that reason over partially observed or hypothetical systems, the CCE operates in fully observed video settings, where the causal process has already unfolded. After observing the complete video, the CCE identifies the actualized, linear sequence of events that led to the outcome—making the task descriptive and post-hoc rather than inferential, and emphasizing interpretability and causal fidelity within the VideoQA framework. However, a key challenge lies in training CCE. In particular, there are no datasets containing reasoning traces for training CCE. To tackle this challenge, we develop an approach to generate causal chains from existing VideoQA datasets efficiently.

Extensive experiments on three large scale datasets demonstrate: **1)** causal chains are promising intermediate representations; **2)** performance improvements across all three datasets; **3)** human studies showed that causal chain-driven video QA enhances explainability and interpretability from multiple perspectives; **4)** the causal chain extractor generalizes well to out-of-domain datasets, highlighting its potential as a reusable causal reasoning engine.

Our main contributions can be summarized as:

- Proposing a structured paradigm for Causal-Why Video Question Answering that leverages natural language causal chains as intermediate representations to enhance reasoning and transparency.
- Introducing a two-stage architecture—Causal Chain Extractor (CCE) and Causal Chain-Driven Answerer (CCDA)—that decouples video understanding from causal inference.
- Introducing a human-in-the-loop framework that uses large language models to propose causal chain drafts, which are subsequently verified and finalized by human annotators, yielding high-quality causal reasoning data.
- Introducing CauCo score, a causality-oriented captioning metric.
- Demonstrating via extensive experiments & carefully designed human studies that the proposed approach outperforms state-of-the-art models while offering significant gains in explainability, user trust, & system debuggability.
- Showing that the CCE generalizes well to out-of-domain datasets, enabling effective causal reasoning across diverse video domains.

2. Related Work

Chain-of-Thought (CoT) is a prompting technique that improves LLMs by decomposing tasks into intermediate reasoning steps. It is effective in arithmetic, logic, and commonsense tasks, especially via few-shot prompting with exemplars [51]. Unlike CoT methods that treat reasoning as emergent, we model reasoning steps—causal chains—as explicit, structured variables. This enables supervision, interpretability, and integration with modular architectures, breaking from the monolithic CoT paradigm.

Causal Reasoning and Structural Causal Models (SCMs). Causality is foundational in scientific reasoning and is formalized in AI through SCMs and do-calculus [40]. Recent work emphasizes causal representation learning for robustness and generalization [44], and SCM-based methods have been used for tasks like debiasing [47] and counterfactual explanations [34]. However, such approaches are rarely applied to VideoQA. We address this gap by integrating SCMs into a modular framework where video-derived causal chains support interpretable answer generation—bringing structured causal reasoning to a domain where temporal complexity often obscures transparency.

Causal Chains and Structured Reasoning in AI. Causal chains provide intuitive, stepwise explanations of how events unfold [35, 38]. In AI, structured representations like causal graphs improve reasoning and interpretability, and prior work has explored causal interactions in videos or intervention-based action recognition [3, 49]. However, these approaches do not use causal chains as intermediates for reasoning. We instead formalize causal chains as explicit bridges between video observations and high-level causal understanding, enabling a principled and interpretable framework for Causal-Why Video QA.

Video Causal Reasoning & Temporal Understand-

ing. Recent work explores causality in video understanding—*e.g.*, structural models for causal interactions and moment retrieval [27, 55]—but not for QA. Other methods select key segments for QA [52] or model object interactions generatively [7], yet their causal reasoning remains implicit or localized. Chen *et al.* [8] introduce event-level causal diagram annotation, but their models explain a final event rather than generate causal chains or use them for QA. In contrast, we extract event-level causal chains and leverage them in a dedicated inference module. Zang *et al.* [57] study video–text causal links, and Su *et al.* [46] generate causal questions from captions, but neither models causal chains or decouples reasoning from answering.

Vision-Language Models (VLMs) & their Limitations in Causality. Recent VLMs (*e.g.*, VideoLLaMA [58], VideoChat2 [26], VILA [30], *etc.*) advance VideoQA & captioning through largescale video–text pretraining, but primarily recognize and describe events rather than explain them. Their monolithic architectures offer limited interpretability & often rely on shallow correlations. Our method complements them by introducing causal chains as structured reasoning intermediates, improving both performance & transparency in Causal-Why Video QA.

3. Causal Chain Construction for SFT

We propose a novel, scalable and efficient methodology to construct causal chains to be used for supervised finetuning (SFT) of our causal chain extractor (Section 4). We generate causal chains that accurately reflect the reasoning behind human-written answers to video-based questions.

Base datasets: since no existing VideoQA dataset includes causal chain annotations to support our approach, we collect causal chains for three challenging causal video QA datasets: **1)** *NextQA* [54], **2)** *CausalVidQA* [23], and **3)** *CausalChaos!* [38]. We refer to these three datasets as the base (or source) datasets. We focus on *Causal-Why QA* in these base datasets. In base datasets, each sample is a paired triplet of {Video ($\mathcal{V}$), Question ($\mathcal{Q}$), Answer ($\mathcal{A}$)}.

Causality defined in the base datasets: the causal-why questions and correct answers in base datasets are human-authored, as a result the notion of cause or causality is well defined in the data itself. All the annotations are rigorously cross-annotator verified, resulting in multihuman agreement on causality. Our goal is to learn to reason through these causal relations. We have provided further details on these three base datasets in the Appendix.

Preliminary I. Human annotators watch the video ($\mathcal{V}$), formulate a question ($\mathcal{Q}$) about it, and write the correct/gold answer ($\mathcal{A}$). The annotations are cross-verified, ensuring multi-annotator agreement. Annotators implicitly use causal chains ($\mathcal{C}$) when writing answers, as illustrated in Figure 2(1). Thus, the detailed QA pairs of our base datasets contain causal chains embedded within them.

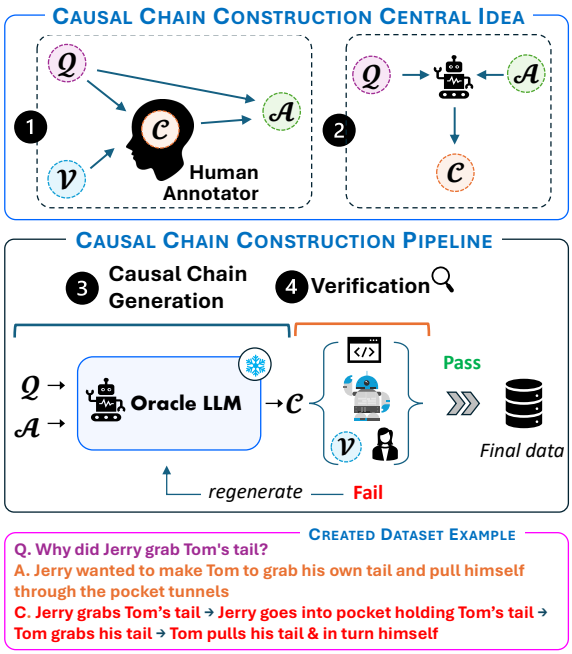

Figure 2. **Causal chain construction for SFT. (1)** Human annotators of base datasets intuitively and implicitly make use of causal chains when writing correct answers. **(2)** We propose to recover these causal chains with the help of LLM using questions and correct gold answers. **(3,4)** Our robust causal chain generation and manual verification and video grounding check pipeline.

Preliminary II. Given the questions and the corresponding correct, gold answers, LLMs have been demonstrated to accurately recover the intermediate reasoning steps or causal chains. Illustrated in Figure 2(2).

Corollary I. Based on preliminaries I and II, we propose that if provided with the questions and the corresponding human written correct, gold answers from the base datasets, then powerful oracle LLMs (*e.g.*, [10, 36, 53]) can reliably recover the reasoning steps or causal chains.

Causal Chain Construction Methodology. Following Corollary I, we propose a robust methodology for causal chain generation and verification in the following.

1. **Causal Chain Draft Generation Process:** We leverage a powerful LLM (GPT4o). We refer to it as Oracle LLM. We prompt the Oracle LLM with question ($\mathcal{Q}$) and the corresponding human written correct, gold answer ($\mathcal{A}$) and ask it to generate a causal chain ($\mathcal{C}$) in natural language in a specified format. Specifically, we instruct the LLM to return causal chains in a structured format: `[Event A] → [Event B] → [Event C]...`[1] Full prompt provided

[1]Note that we do not train the LLMs to generate causal chains; rather, directly capitalize on their inherent knowledge, commonsense reasoning & synthesis capabilities to derive them.

in the Appendix. Illustrated in Figure 2(3). Oracle LLM is not used as end-task solvers but as structured reasoning components within a rigorously controlled pipeline. Their role is analogous to that of annotator assistants—providing drafts that undergo multi-layer human validation to ensure reliability and reproducibility.

2. **Causal Chain Verification Process:** Illustrated in Figure 2(4). We ensure that the generated causal chains are of high quality and accurately capture the reasoning steps using rigorous Quality Checks consisting of:

(a) ***Programmatic Validation:*** Programs check generated causal chains for structure, format, completeness, and length ($\leq$10 events).

(b) ***Cross-LLM Verification:*** Verifier LLM is provided with question ($\mathcal{Q}$), correct gold answer ($\mathcal{A}$) and the generated causal chain ($\mathcal{C}$) Generated causal chains are independently reviewed and verified by second LLM to assess the chain's logical coherence, relevance, and consistency with $\mathcal{Q}$ and $\mathcal{A}$ pair. For this, we use a powerful LLM, but different than Oracle LLM (GPT4o) to avoid LLM circularity bias. Specifically, we use Gemini 2.5 as the verifier LLM.

(c) ***Manual Correction & Verification against $\mathcal{V}$:*** Human verifiers receive the video $\mathcal{V}$, question $\mathcal{Q}$, and correct gold answer $\mathcal{A}$ to ensure proper grounding of the causal chain in the video content. Chains passing the first two stages are checked for logical coherence, relevance, and consistency with $\mathcal{Q}$, $\mathcal{A}$, and $\mathcal{V}$. Verifiers—computer science graduates familiar with the task—may add missing details to causal chains and flag hallucinations (*e.g.*, events not observable or implied in the video). Chains with hallucinations or other failures are regenerated until they pass all checks.

Only the chains that have passed all the checks are considered for the final dataset. Finally, 1000 samples belonging to each base dataset are manually reviewed by the authors. Over 95% of the chains passed the author verification, ensuring accurate causal chain annotations. Our method produced reliable, accurate causal chains grounded in the video context. In total, we constructed human-verified causal chains for 46,024 samples across three datasets. Further stats provided in the Appendix.

4. Approach

4.1. Overview and Motivation

Conceptually, our work is grounded in the principles of Structural Causal Models (SCMs) [40] and Chain-of-Thought (CoT) [51] reasoning. We decompose the VideoQA task into two explicit stages: **1)** causal chain extraction; and **2)** causal chain-driven answering. Please refer to Figure 1. From an SCM perspective, we introduce causal chains as structured intermediate representations. While many existing VideoQA approaches can be abstracted as an undifferentiated model where video ($\mathcal{V}$) and question ($\mathcal{Q}$) jointly influence the answer ($\mathcal{A}$) as ($\mathcal{V} \rightarrow \mathcal{A} \leftarrow \mathcal{Q}$), we propose a *structured reasoning via a Causal Reasoning Trace* ($\mathcal{C}$), yielding the following SCM: $\mathcal{V} \rightarrow \mathcal{C} \rightarrow \mathcal{A}, \mathcal{Q} \rightarrow \mathcal{C} \rightarrow \mathcal{A}$. Then, we *factorize* this SCM into: $\mathcal{V} \rightarrow \mathcal{C} \leftarrow \mathcal{Q}$ (this becomes our *Causal Chain Extractor*—Figure 1(3)) and $\mathcal{Q} \rightarrow \mathcal{A} \leftarrow \mathcal{C}$ (this becomes our *Causal Chain-driven Answerer*—Figure 1(4)). Causal Chain Extractor module would first extract $\mathcal{C}$, and pass it to the Causal Chain-Driven Answerer module.

We assume access to fully observed videos, where the causal process leading to an outcome manifests as a temporally ordered sequence of events—a single realized traversal through the underlying causal graph. Once this process is complete, the relevant causal information resides entirely within this factual path. Thus, for Causal-Why VideoQA, the causal chain can be modeled as a linear sequence of observed cause–effect events, providing a complete & interpretable representation of the reasoning process. Detailed treatment of the proposed use of linear causal chains is provided in the Appendix. The structured factorization aligns with CoT principles by making reasoning steps explicit, & reflects SCM modularity by separating causal understanding from answering. As a result, our approach enables:

- **Focused processing:** Unlike existing VideoQA methods, our structured approach enables focused processing, with each stage's output passed to the next. Processing smaller chunks reduces the risk of missing reasoning steps or making incorrect inferences. It is inspired by the Chain-of-Thought philosophy, adapted into a structured model.
- **Improved video understanding:** Modeling cause-and-effect relationships explicitly in causal chains is a dense prediction task. With detailed supervision, vision models learn to better capture rich video content. In contrast, typical VideoQA models are trained with a single label, which may be insufficient to capture the video's complexity [6].
- **Enhanced explainability:** By generating human-readable causal chains as intermediate outputs, our model improves both explainability & interpretability, making the reasoning process more transparent.

4.2. Model

Thus far, we saw that our approach factorizes VideoQA problem into modules: causal chain extraction and causal chain-driven answerer. Now, we explain these individual modules and their training/inference in the following.

4.2.1. Causal Chain Extractor (CCE)

The objective of CCE module is to explicitly capture the cause-and-effect relationships or the reasoning steps from videos, conditioned on the questions, and express them in a detailed yet concise form. The CCE extracts the causal chain after the entire video has been observed, when the full causal process and its outcome are known. At this stage, the causal structure is realized as a single factual sequence of

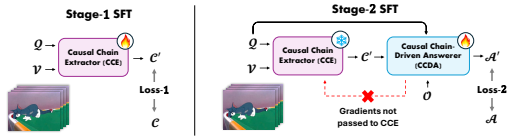

Figure 3. **Stage-wise training of our model.**

events leading to the answer. The extractor therefore identifies this linear causal path directly, without reasoning over counterfactual or hypothetical alternatives. This formulation treats causal chain extraction as a descriptive recognition task, emphasizing interpretability, fidelity, and robustness within the VideoQA paradigm. Generating causal chains from videos conditioned on questions is a complex reasoning task. We leverage the representational capacity of foundation models. However, generating causal chains is non-trivial and beyond existing vision-language foundation models. Although foundation models exhibit strong performance across vision-language benchmarks, they lack structured causal supervision, limiting causal chain generation. To address this limitation, we introduce a novel dataset of video–question–causal chain triplets (Section 3). This dataset enables supervised fine-tuning (SFT) of pretrained vision-language models for structured causal inference. This trains the model to capture event-level dependencies and generate human-interpretable causal chains grounded in visual and linguistic context. Formally, given a video $\mathcal{V}$ and a corresponding question $\mathcal{Q}$, the model learns a mapping: $f_{CCE} : (\mathcal{V}, \mathcal{Q}) \rightarrow \mathcal{C}$; where $\mathcal{C}$ denotes the generated causal chain and f_{CCE} represents the fine-tuned vision-language foundation model. Further model implementation details provided in Appendix.

4.2.2. Causal Chain-Driven Answerer (CCDA)

The objective of CCDA module is to select the correct answer from candidate answers, based on the questions and the causal chains. Existing large language models have shown strong performance on a variety of language-based tasks. We propose leveraging language models to implement the CCDA. However, processing causal chains using language models is non-trivial and mostly not covered in their training suites, largely due to lack of explicit causal chains or step-by-step annotations of causal reasoning. Towards that end, we resort to supervised finetuning of CCDA. Specifically, CCDA receives the extracted causal chain, question, and candidate answer options (typically four to five in existing datasets) and is prompted to select the correct answer. Notably, the Causal Chain Extractor (CCE) remains frozen at this stage. Formally, given a question $\mathcal{Q}$, the corresponding causal chain $\mathcal{C}$, and answer options $\mathcal{O}$ the model learns a function: $f_{CCDA} : (\mathcal{Q}, \mathcal{C}, \mathcal{O}) \rightarrow \mathcal{A}$; where $\mathcal{A}$ denotes the selected answer option and f_{CCDA} represents a finetuned language model.

4.2.3. Clean Separation with Stage-wise Training

Problem with end-to-end (E2E) training. E2E VQA models optimize all components—from feature extraction to answer prediction—using final answer loss. This causes gradient leakage, letting error signals flow through the entire pipeline. As a result, models may learn "shortcuts" that boost final answer accuracy but harm causal reasoning.

Our solution—stagewise training. Our two-stage, stagewise training prevents this by introducing a clear boundary:

- *CCE Training:* The CCE is trained independently to match the ground-truth causal chains (C), using a loss defined on C that enforces accurate causal grounding.
- *CCDA Training:* The CCDA is trained on the CCE's output, with the CCE's weights typically frozen or its predictions treated as fixed during this stage.

By freezing the CCE, no gradients can pass backward from the CCDA's answering loss to the CCE's internal weights. This ensures that the CCE's learned causal reasoning logic remains pure and untainted by the pressure to maximize the answer score, thus preserving the integrity, robustness, and causal semantics of the intermediate representation.

4.2.4. Inference

During inference, the CCE predicts a causal chain from the video and question *without* using ground-truth chains. The CCDA then uses this chain, the question, and answer options to select the correct answer.

4.3. CauCo: A Causal Coherence Metric

We introduce the CauCo score to evaluate how well generated causal chains model causality—an aspect regular captioning metrics miss. CauCo quantifies causal consistency and provides causal guarantees in open-world settings by measuring whether events are logically linked by cause and effect. Following the LLM-as-verifier approach, we SFT an LLM to judge whether a chain is causally coherent. The evaluator outputs "True" or "False." For SFT, we construct positive and negative samples: positive ones are correct causal chains, and negatives are created by perturbing them using six strategies—*actor swapping*, *event negation*, *event removal*, *event order reversal*, *semantic modification*, and *chain shuffling*.

5. Experiments

This section presents experiments validating our hypotheses. We first establish an upper bound for causal VideoQA using our method, then compare it to state-of-the-art approaches, and finally assess explainability through human studies. The experimental setup is outlined below.

Implementation details. We use PyTorch [39] to implement models. Noting the strong performance of VILA 1.5 & LLaMA on vision and language tasks, we adopt them as representative models in implementing our CCE and

Dataset	NextQA	CVQA	CausalChaos!	Overall
CCDA Accuracy	99.70	99.85	98.65	99.40

Table 1. **Experimental results (Accuracy in %) for answering based on ground truth causal chains.**

CCDA, respectively. Note that our approach is not designed for or limited to any models; practitioners may use models of their choice such as [9, 26, 30, 58]. Our CCE is based on VILA-3B [30]; CCDA is LLaMA-3.1-8B [17] version. Further implementation details are provided in Appendix.

Datasets. We conduct experiments on the datasets as discussed in Section 3.

Tasks. We evaluate on a multiple-choice QA task with five answer options per question, only one of which is correct.

Performance metric. Following prior work [23, 38, 54], we use Accuracy as the performance metric. For causal chain quality, we use our CauCo score and captioning metrics: BLEU [37], Meteor [5], ROUGE [29], SPICE [1].

5.1. Experimental Upper bound on Performance

We first investigate whether causal chains can serve as effective intermediate representations. To this end, we use groundtruth causal chains instead of predicted ones, *training & testing* the Causal Chain-Driven Answerer on these annotations. This setting simulates an ideal scenario where causal chain generation is perfectly accurate. Thus, this experiment aims to determine the model's performance upper bound when provided with flawless causal chains.

The results, summarized in Table 1, show near-perfect accuracy, even surpassing human performance. This finding suggests that using causal chains as intermediate representations is a promising paradigm worth further exploration.

5.2. Ablation Studies on the Role of Causal Chains

To test whether CCDA's improvements arise from genuine causal reasoning rather than surface-level context enrichment, we conducted ablation studies that systematically degrade the quality of causal chains. **Study-I:** We semantically perturbed the causal chains such that contextual information was preserved but causal relations were disrupted. We observed a sharp decline in QA accuracy once causal information was perturbed, despite the context remaining intact (Figure 4(1)). **Study-II:** We progressively masked links and measured QA performance; accuracy degraded monotonically with increased perturbation (Figure 4(2)). Causal-chain quality strongly correlated with CCDA accuracy (r=0.97). These results empirically validate that CCDA's reasoning depends on causal-chain integrity, with improvements driven by causal reasoning rather than spurious correlations. Qualitative examples (Figure 4(3,4)) show how chain perturbations intuitively alter CCDA's answers.

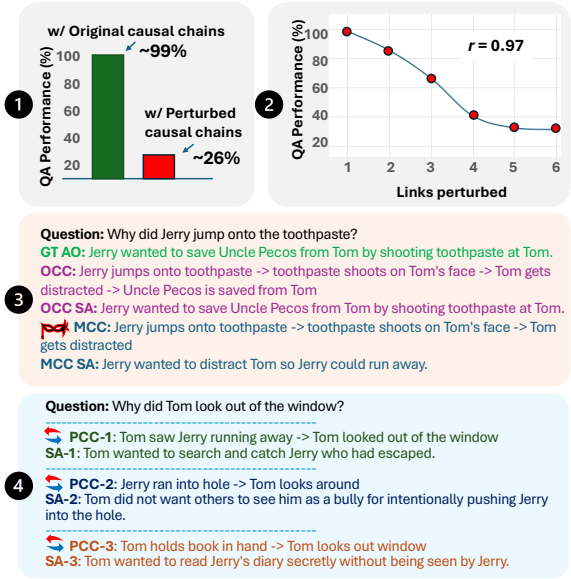

Figure 4. **Causal chain ablation study. (1)** Study-I: QA accuracy drops by 73% when chains are perturbed. **(2)** Study-II: drop in QA is correlated to amount of perturbation. **(3,4)** Qualitative example. OCC: original causal chains, MCC: masked chains, SA: selected answer, PCC: perturbed chains. SA changes intuitively as chains are perturbed. *Please zoom-in.*

Model	NeXT-QA [54]	CausalVidQA [23]	CausalChaos!QA [38]	Avg	WtAvg
Traditional VideoQA approaches					
BlindQA [2]	28.38	59.46	13.07	33.64	44.87
EVQA [2]	42.31	60.95	13.48	38.91	50.62
CoMem [15]	46.15	62.79	13.88	40.94	53.05
HME [12]	46.52	61.45	14.02	40.66	52.43
HCRN [21]	47.00	61.61	17.00	41.86	52.93
HGA [20]	47.00	63.51	15.36	42.00	53.88
Semi-traditional approaches					
MIST [14]	54.79	72.41	44.88	57.36	64.04
Causal modeling approaches					
VCSR [52]	53.00	65.41	-	-	-
MCR+HCRN [57]	49.20	66.00	-	-	-
VLMs/Multimodal foundation approaches					
BLIP-2 [25]	45.00	62.00	23.32	43.44	53.00
VideoLLaMA [58]	42.00	31.00	11.73	28.24	33.31
VideoChat2 [26]	60.00	46.00	15.36	40.45	48.50
GPT4o [36]	70.00	52.00	48.17	56.72	58.00
VILA 1.5-3B [30]	60.23	72.11	62.80	65.05	67.20
DeepSeek-VL2 [53]	51.55	57.08	17.12	41.92	51.95
QwenVL 2.5-7B [4]	70.75	76.22	31.54	59.50	70.73
Ours	63.95	76.18	67.65	**69.26**	**71.22**

Table 2. **Performance evaluation with SOTA methods.** WtAvg: Weighted average.

5.3. Performance Comparison with SOTA

Baselines. Following prior work [23, 38, 54], we compare against a wide range of models: **1)** *traditional approaches* [2, 12, 15, 20, 21]; **2)** *causal approaches* [52, 57]; and **3)** *SOTA VLMs/Multimodal foundation approaches* [4, 14, 25, 26, 30, 36, 53, 58], which excel on vision-language tasks.

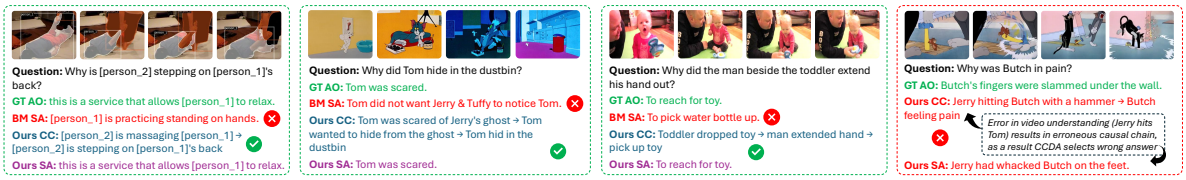

Figure 5. **Qualitative results.** GT AO: Groundtruth Answer Option; BM: Baseline Model; SA: Selected Answer; CC: Causal Chain. Only a few frames per video are shown. Green and red boxes indicate success and failure cases. In the first example, actor masks come from the CausalVidQA dataset, which includes reference-based QA. *Please zoom-in.*

Model	B1	B2	B3	B4	M	R	S	CCS
QwenVL2.5-3B Oneshot	0.35	0.23	0.16	0.11	0.54	0.36	0.40	0.75
Ours	**0.63**	**0.47**	**0.36**	**0.28**	**0.61**	**0.50**	**0.52**	**0.89**

Table 3. **Causal chain generation performance results.**

Model	Explainability	Trustworthy	Human Preferred
Blackbox	01.33	01.33	14.81
Ours	**69.33**	**62.67**	**85.18**
No Preference	29.33	36.00	n/a

Table 4. **Human study results (%) along various axes.**

Results. Table 2 reveals that traditional VideoQA models lag behind VLMs. MIST, a smaller model, performs well—likely from its long-term video understanding and use of CLIP features with spatiotemporal attention. Our method further outperforms prior causal models such as MCR+HCRN [57], the closest to our work, while remaining simpler and producing interpretable causal chains, unlike latent-variable or counterfactual methods. Next-generation VLMs such as VILA [30] outperform GPT-4o on causal VideoQA. Building on VILA, our framework surpasses prior models through improved visual grounding via causal chain extraction and explicit CCDA reasoning. In contrast, VLMs perform grounding and reasoning jointly, often overlooking visual evidence and relying on plausibility biases. Our structured model guides the extractor to capture cause–effect relations in videos—a dense prediction task enhanced by fine-grained supervision. Typical VideoQA models use a single-label objective, leading to shallow learning, whereas our approach isolates comprehension from bias, allowing the CCDA to focus on selecting answers from extracted causal chains. The CCDA made fewer errors, though the extractor occasionally misinterpreted object roles or actions, producing flawed chains and reducing accuracy (Figure 5). Based on our analysis, future work should focus on **1)** role or relationship modeling and **2)** situation understanding. We also compare our model's causal chains with those from QwenVL2.5, a representative SOTA VLM (Table 3, averaged across datasets). Our model performs significantly better, suggesting that causality is often overlooked in multimodal understanding [32]—a gap our work addresses. We believe models like ours also have the potential to serve as reasoning engines, which should be explored by future work.

5.4. Human Studies

Explainable systems benefit two user groups: researchers/ system designers & consumers. We design four studies to evaluate our explainable system against black-box models across multiple criteria. Since most SOTA VLMs are black boxes, we use VILA [30] as a representative model for its strong performance & lack of explicit explanations. To ensure reliability, we conducted user studies with six participants, exceeding the sample size used in prior work [38]. To minimize bias: 1) no study participants were involved in the project; 2) Sample order randomization & blinding (independent, anonymized tests) was used for all participants. Further details provided in Appendix.

5.4.1. Study I: Explainability

In this study, participants are presented with 50 questions and models' outputs, including final answers and, when available, intermediate explanations in the form of causal chains. They rate their understanding of the explanation,*i.e.*, causal chains, regardless of the correctness of the final answer. To avoid the ambiguity of arbitrary scales (*e.g.*, 1–5), we use a comparative evaluation: participants choose between: **1)** *System A*—no explanation (*BlackBox model*); **2)** *System B—Our Model* with causal chain explanations; or **3)** *No Preference*—indicating no added value from the causal chains as explanations. Table 4 shows participants found causal chains useful explanations in over 69% of cases, demonstrating their overall effectiveness. Nonetheless, in ∼29% of cases, the explanations were not considered helpful, suggesting opportunities for further refinement.

5.4.2. Study II: Trustworthiness

In this study, participants are presented with 50 questions and models' outputs, including final answers and, when available, intermediate explanations in the form of causal chains. The goal is to evaluate users' trust in the systems based on the explanations and the perceived correctness of the answers, without access to ground-truth labels. Partic-

Model	VLM	LLM	Cannot Tell
Blackbox	15.00	15.00	70.00
Ours	**48.33**	**38.33**	**13.33**

Table 5. **Utility from system debugging perspective.**

ipants select one of the following options: **1)** *System A*—trust the *BlackBox* model's answer; **2)** *System B*—trust *Our Model's* answer with its causal chain explanation; or **3)** *No Preference*—no clear preference between the two. Table 4 summarizes that in over 62% of cases, participants expressed significantly greater trust in our model. However, trust declined when the model's predictions conflicted with participants' expectations. We anticipate that trust will increase as the accuracy of our approach improves.

5.4.3. Study III: Human Preference

In this study, we analyze human preference between the two systems: **1)** *BlackBox model*; or **2)** *Our Model*. To ensure a fair comparison, we consider only successful predictions (*i.e.*, those matching groundtruth answers) from both models. Human participants evaluate 50 samples, selecting their preferred model. Participants are instructed to assume *they are using an AI system designed to study human behavior and explain the "why" behind people's actions in the videos*. The results of this study are presented in Table 4. We found that in over 85% of cases, participants strongly preferred a system that provides explanations, like our approach, to support its decisions.

5.4.4. Study IV: Utility in System Debugging

Researchers & system designers seek to improve models via systematic debugging. Identifying limitations is the first step; intermediate outputs such as causal chains alongside final predictions provide useful insights. We tested this by examining failure cases of our explainable and black-box models. While this assumes known errors, it reflects real settings where designers work with labeled data. Researchers are then asked to identify the source of the model's failure: **1)** visual perception, **2)** language and reasoning, or **3)** Cannot Tell. The third option indicates uncertainty, while choosing the first two suggests a clearer understanding of the system and supports more confident diagnosis. Six researchers with expertise in computer vision, VLMs, & LLMs each analyze 20 failure cases from both systems, identifying fault locations. For our model, they view the video, question, causal chain explanations, & final predictions; for the black-box model, only the video, question, and final predictions are provided. Participants are instructed to use all available information in their assessments. Table 5 shows that "Cannot Tell" was selected far less often for our model than for BlackBox. Researchers attributed ∼48% of failures to the chain extractor and ∼38% to the chain-driven answerer, indicating improved debuggability from causal chains.

Model	B1	B2	B3	B4	S	R	M	CCS
Baseline	0.19	0.10	0.06	0.04	0.35	0.29	0.37	0.42
Ours	**0.35**	**0.22**	**0.13**	**0.08**	**0.39**	**0.37**	**0.61**	**0.84**

Table 6. **OOD chain generation performance evaluation.**

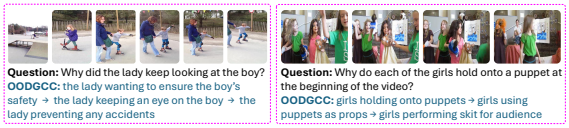

Figure 6. **Qualitative examples of Out-of-Domain generated causal chains (OODGCC).** *Please zoom-in.*

5.5. Out-Of-Domain Chain Generation

To test the generalizability of CCE, we train it on a cartoon-based dataset (CausalChaos!) and evaluate it on an out-of-distribution real-world dataset (NextQA), making the task intentionally challenging. Quality of generated chains is evaluated in terms of BLEU-1–4 (B), METEOR (M), ROUGE (R), SPICE (S), our causal coherence score CCS, w.r.t. groundtruth causal chains (Section 3). A zero-shot VILA 1.5 serves as the baseline; our extractor also uses VILA 1.5 for a fair comparison. Table 6 shows that our model significantly outperforms the baseline in cross-dataset causal chain generation—suggesting that the reasoning patterns that our causal chain extractor learns are robust and transferable, highlighting its potential as a reusable causal reasoning engine. Qualitative results in Figure 6.

6. Conclusion

We shift paradigms and introduce a principled, structured approach for Causal-Why VideoQA task, which decouples video understanding, reasoning and answer generation. These modules use natural language causal chains as intermediate representations. Our approach enforces analyzing entire videos and explicitly reason to determine the cause behind the actions, instead of shortcut-based answering. At the core of our method lies novel integration of SCMs and CoT. Our approach We show how to efficiently construct training data to develop such models. Thorough experimentation demonstrates that our approach improves explainability, performance on VideoQA task, and have the potential to serve as reusable causal reasoning engines.

Acknowledgements. This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award# NRF-NRFF14-2022-0001). This research is also supported by funding allocation to B.F. by the Agency for Science, Technology and Research (A*STAR) under its SERC Central Research Fund (CRF), as well as its Centre for Frontier AI Research (CFAR). We would also like to thank Dhruv Verma and Elston Tan.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 6

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 6, 19

[3] Mustafa Ayazoglu, Burak Yilmaz, Mario Sznaier, and Octavia Camps. Finding causal interactions in video sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3575–3582, 2013. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[6] Jing Bi, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Guangyu Sun, Jinxi He, et al. Why reasoning matters? a survey of advancements in multimodal reasoning (v1). *arXiv preprint arXiv:2504.03151*, 2025. 4

[7] Guangyi Chen, Yuke Li, Xiao Liu, Zijian Li, Eman Al Suradi, Donglai Wei, and Kun Zhang. Llcp: Learning latent causal processes for reasoning-based video question answer. In *The Twelfth International Conference on Learning Representations*, 2023. 3

[8] Tieyuan Chen, Huabin Liu, Tianyao He, Yihang Chen, Xiao Ma, Cheng Zhong, Yang Zhang, Yingxue Wang, Hui Lin, Weiyao Lin, et al. Mecd: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37:92554–92580, 2025. 3

[9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3

[11] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 16

[12] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 6, 19

[13] 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. 1

[14] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783, 2023. 6, 19

[15] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 6, 19

[16] Alison Gopnik, Clark Glymour, David M Sobel, Laura E Schulz, Tamar Kushnir, and David Danks. A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3, 2004. 2

[17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[18] Norwood Russell Hanson. Causal chains. *Mind*, 64(255): 289–311, 1955. 2

[19] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 19

[20] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11109–11116, 2020. 6, 19

[21] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 6, 19

[22] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 20

[23] 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 (CVPR)*, 2022. 1, 3, 6, 17

[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 (CVPR)*, 2022. 20

[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 *ICML*, 2023. 6

[26] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. 2024. 3, 6, 19

[27] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33:9180–9192, 2020. 3

[28] Jianxin Liang, Xiaojun Meng, Huishuai Zhang, Yueqian Wang, Jiansheng Wei, and Dongyan Zhao. Reasvqa: Advancing videoqa with imperfect reasoning process. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1696–1709, 2025. 21

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[30] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 3, 6, 7, 19

[31] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020. 20

[32] Jing Ma. Causal inference with large language model: A survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5886–5898, 2025. 7

[33] David Mogrovejo and Thamar Solorio. Question-instructed visual descriptions for zero-shot video answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9329–9339, 2024. 21

[34] Tanmayee Narendra, Anush Sankaran, Deepak Vijaykeerthy, and Senthil Mani. Explaining deep learning models using causal inference. *arXiv preprint arXiv:1811.04376*, 2018. 2

[35] Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. R^ 3-vqa:" read the room" by video social reasoning. *arXiv preprint arXiv:2505.04147*, 2025. 2

[36] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. [Online; accessed 31-May-2024]. 3, 6

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[38] Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1, 2, 3, 6, 7, 17, 18, 20

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[40] Judea Pearl. *Causality*. Cambridge university press, 2009. 2, 4

[41] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018. 2

[42] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4932–4942, 2019. 21

[43] Ishaan Singh Rawal, Alexander Matyasko, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. Dissecting multimodality in videoqa transformer models by impairing modality fusion. In *International Conference on Machine Learning*, pages 42213–42244. PMLR, 2024. 1

[44] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 2

[45] Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8523–8533, 2025. 21

[46] Hung-Ting Su, Yulei Niu, Xudong Lin, Winston H Hsu, and Shih-Fu Chang. Language models are causal knowledge extractors for zero-shot video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4951–4960, 2023. 3

[47] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 2

[48] Tom Trabasso and Paul Van Den Broek. Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630, 1985. 2

[49] Yuqing Wang, Lei Meng, Haokai Ma, Yuqing Wang, Haibei Huang, and Xiangxu Meng. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3936–3944, 2024. 2

[50] Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 92–101, 2024. 21

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2, 4

[52] Yushen Wei, Yang Liu, Hong Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 377–386, 2023. 1, 3, 6

[53] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 3, 6

[54] 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. 1, 3, 6, 17, 20

[55] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1–10, 2021. 3

[56] Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21655–21666, 2025. 21

[57] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19027–19036, 2023. 3, 6, 7

[58] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics. 3, 6, 19

[59] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision. *arXiv preprint arXiv:2407.06189*, 2024. 21