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

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

Appendix

Contents

7. Discussion on Causal Chain structure

In this section, we elaborate on the motivation for adopting linear causal chains as intermediate representations in our Causal-Why VideoQA approach. Although the underlying causal structure of the world is often complex, involving branching, feedback, and multi-agent interactions, a fully observed video corresponds to a single factual realization of this structure—a specific trajectory of events that actually occurred. This realized causal process naturally takes the form of a linear, temporally ordered sequence of events. In the following, we discuss how this linearity arises, why it is conceptually justified and experimentally validated, and how it provides a principled and practical foundation for modeling causality in video-based reasoning tasks.

7.1. Causal Graphs and Realized Causal Chains

Causal graphs, such as those represented in structural causal models (SCMs) or dynamic Bayesian networks, describe the space of all possible causal interactions that could occur within a system. These graphs can include cycles, parallel pathways, or multiple interacting agents. However, in a fully observed video, the viewer does not witness the entire causal graph but rather a single factual instantiation of it—a concrete, temporally unfolding sequence of events that actually takes place. Even when the underlying causal graph is cyclic or contains alternative pathways, a single realized execution of the process constitutes a traversal through that graph. This traversal can be expressed as a temporally ordered chain of events: $e_1 \rightarrow e_2 \rightarrow ... \rightarrow e_T$, where each event corresponds to a distinct moment or causal transition within the observed sequence. The Causal Chain Extractor (CCE) in our framework captures precisely this notion: it models not the full causal topology, but the specific event sequence manifested in the video.

7.2. On Linearity of Causal Chains

The linearity of causal chains arises fundamentally from the linear nature of time in observational media. Every video imposes an intrinsic temporal ordering on events—each event occurs before, after, or simultaneously with another. This sequential structure ensures that, regardless of the complexity of the underlying causal mechanisms, the observed process can always be expressed as a linear progression over time. Unrolling the system temporally yields a sequence $e(t_1) \rightarrow e(t_2) \rightarrow e(t_3) \rightarrow ...$, analogous to how recurrent neural networks or dynamic causal models unfold feedback mechanisms into linear time-indexed representations. Thus, linearity in our framework is not an artificial constraint, but a natural consequence of the temporal structure inherent to video data.

7.3. Linear Causal Chains Accomodate Complex Causality

Although videos unfold continuously over time, events within them do not always occur in a strictly sequential fashion. Multiple causes may arise simultaneously, or a cause and its effect may partially overlap in duration. Our formulation of a linear causal chain therefore does not impose a discretization of the video into rigid, non-overlapping time steps. Instead, it encodes the logical progression of causation, capturing how influences propagate, whether or not the underlying events are temporally distinct. In this sense, the linear chain represents the ordering of causal influence rather than the literal ordering of clock-time frames.

To accommodate simultaneity, a causal chain may include composite events that group together several co-occurring actions or states. A node such as $e_k = \{A, B\}$ denotes that events A and B occur jointly or lack a meaningful temporal separation with respect to the causal question. The subsequent step in the chain, $e_k \rightarrow e_{k+1}$, then captures the next causal transition, even though the preceding node aggregates multiple parallel or interacting components. In this way, temporal linearization is conceptual

rather than strictly chronological: it describes what leads to what without asserting that the corresponding events occur one after another in physical time.

This flexible representation allows linear causal chains to encompass a wide range of temporal relationships. Simultaneous agent actions such as "Butch and Tom both lunge at Jerry," joint causes like "both characters pull the lever at the same moment," or causally inseparable event clusters such as "the vase shatters as it hits the floor" can all be expressed within the same linear structure. Even micro-overlaps that are indistinguishable at the frame level are naturally accommodated. Consequently, the linear-chain formulation preserves the causal intelligibility and explanatory clarity of event sequences while remaining faithful to the nuanced temporal structure of real video phenomena.

7.4. Advantages of the Linear-Chain Perspective

Our adoption of a linear-chain perspective in causal reasoning provides several distinctive advantages over existing approaches in Causal-Why VideoQA. While previous methods typically rely on unstructured textual rationales, spatial attention maps, or implicit latent representations, these formulations fail to capture the explicit ordering and structure of causal reasoning. By contrast, representing causality as a linear chain of temporally (or logically) ordered events yields a unified, interpretable, and computationally tractable framework that aligns closely with how humans naturally perceive and describe causal processes. Particularly, a linear-chain perspective in causal reasoning provides the following advantages:

1. The linear-chain formulation offers a unifying representation for complex causal phenomena. Real-world events often involve feedback, branching, multiple agents, and parallel actions that complicate traditional graph-based or latent models. The linear-chain approach collapses this complexity into a single, temporally coherent representation that reflects the realized causal trajectory within the observed video. This conversion is not merely a simplification but a principled abstraction: any complex causal graph, when instantiated in time, becomes a single realized path that can be represented as a chain. Consequently, linear causal chains provide a consistent framework for representing diverse causal configurations within a common formalism.
2. Linear causal chains produce explicit, human-interpretable explanations that bridge the gap between model reasoning and human commonsense understanding. Existing VideoQA methods often justify answers using attention distributions or post-hoc rationales that are either opaque or only loosely related to causal dependencies. In contrast, a causal chain expresses reasoning as a stepwise progression of cause and effect, offering a transparent account of how the model infers the answer. This interpretability makes it possible to inspect, verify, and communicate the model's reasoning in human-readable terms, thereby enhancing trust and accountability.
3. The linear-chain assumption imposes structural constraints that serve as a form of regularization during training and inference. Traditional attention-based systems can easily overfit to spurious correlations or irrelevant visual cues, as they are underconstrained in terms of causal ordering. The requirement that reasoning follow a coherent causal chain restricts the solution space, encouraging the model to prioritize temporally and causally consistent evidence. This constraint helps mitigate common issues such as reasoning drift and hallucinated explanations, leading to more grounded and faithful responses.
4. The linear-chain design enables modularization of the reasoning process into two tractable subproblems: causal chain extraction (CCE) and causal chain-driven answering (CCDA). This decomposition allows the system to separately learn to identify causally relevant events and to reason over them, thereby improving both robustness and interpretability. Unlike end-to-end architectures that conflate perception and reasoning, this modular design promotes clearer error diagnosis and facilitates targeted improvements to each stage.
5. Linear causal chains contribute to efficient learning and improved sample efficiency. Because chains summarize only the causally relevant portions of the video, they provide a compact intermediate representation that filters out irrelevant background content. This focused supervision signal allows the model to generalize from fewer examples while maintaining causal coherence. Compared to approaches that require full graph annotations or rely on free-form text rationales, linear chains are easier to annotate, verify, and scale, making them practically advantageous for dataset construction and model training.
6. The linear-chain perspective serves as a practical bridge between visual evidence and symbolic reasoning. Full causal graphs are expressive but computationally intractable for long, high-dimensional video data, while textual rationales lack structural precision. Linear chains occupy an intermediate space: they are structured enough to support formal reasoning, yet simple enough to be generated and interpreted by neural models and large language models (LLMs). Importantly, this formulation aligns naturally with the stepwise "chain-of-thought" reasoning patterns that LLMs exhibit, enabling seamless integration between visual grounding and language-based inference.

7.5. Empirical Validation of the Linear Causal Chain Representation

The empirical upper bound—computed using oracle causal chains derived from human-verified ground-truth causal chains (main paper Section 3)—achieves near-perfect performance (main paper, Experiment 5.1, Table 1). This demonstrates that linear causal chains are an expressive and sufficient representation for capturing the causal dynamics required in Causal-Why VideoQA. Although the linear format is structurally minimalist, the upper-bound results provide strong evidence that it does not restrict the causal reasoning capacity of the system in practice.

Importantly, our use of a linear chain does not oversimplify the true causal structure of the underlying environment. The chain encodes a single realized causal trajectory in the observed video—not the full causal graph—which is the level of abstraction relevant for answering factual, video-grounded causal questions. Moreover, composite events (*e.g.*, $e_k = \{A, B\}$) allow the representation to capture simultaneous, joint, or causally inseparable events, ensuring that complex interactions are not lost during linearization.

In summary, the linear-chain perspective offers both conceptual and empirical/practical advantages. It unifies diverse causal phenomena under a single temporal representation, produces interpretable and verifiable reasoning traces, regularizes learning through causal constraints, supports modular system design, enhances computational efficiency, and provides a bridge between neural and symbolic reasoning paradigms. These advantages collectively underscore why linear causal chains represent a valuable and effective abstraction for causal reasoning in VideoQA.

8. Discussion: Toward a Learnable Causal Reasoning System

A central contribution of our work is the introduction of a learnable causal reasoning system for video-based causal-why question answering. Unlike traditional VideoQA pipelines that fuse perception, reasoning, and answer prediction within a monolithic architecture, our modular framework explicitly learns a structured causal reasoning process. At its core lies the Causal Chain Extractor (CCE), a supervised model that acquires the ability to produce event-level causal explanations from video–question input pairs. The CCE is not a rule-based or templated mechanism; rather, it is trained on human-verified causal chains, enabling it to internalize the patterns and semantics of cause–effect relationships across diverse video scenarios. In doing so, the CCE functions as an end-to-end trainable causal inference module operating in fully observed environments.

8.1. Learning Causal Reasoning from Data

The learnability of the system stems from the following three properties:

1. Causal chains are treated as explicit intermediate supervision. Instead of expecting the model to implicitly "discover" causal structure from answer labels alone, we provide detailed, human-grounded causal reasoning traces, allowing the model to learn how real-world cause–effect sequences manifest in videos. This transforms causal reasoning from an emergent byproduct of multimodal embeddings into a directly optimized capability.
2. The causal chain extractor (CCE) is trained to map complex visual sequences and natural-language questions into multi-step causal chains—an ability that requires understanding temporal dependencies, tracking agency, and interpreting intentions. These tasks themselves constitute a form of machine-learned causal reasoning, but crucially, the system learns this capability through data-driven generalization rather than handcrafted causal graphs.
3. The CCDA (Causal Chain-Driven Answerer) learns to interpret causal chains as explicit reasoning evidence. Its supervised finetuning allows it to internalize causal reasoning patterns (*e.g.*, actor continuity, event progression, causal sufficiency) and use them to select answers. The success of CCDA when fed ground-truth chains demonstrates the robustness of this learned reasoning paradigm—achieving near-perfect performance across datasets.

8.2. Interpretability and Modular Causal Semantics

An important aspect of being a "causal reasoning system" is not only producing correct answers but doing so through transparent causal semantics. The CCE generates natural-language causal chains that function as interpretable reasoning traces. These traces expose the model's internal causal beliefs about the video, allowing users to understand why a prediction was made. This supports both user-facing explainability and model debugging: human evaluators can often diagnose which component—perception or reasoning—was responsible for errors.

The modular design ($\{\mathcal{V}, \mathcal{Q}\} \rightarrow \mathcal{C}$ and $\{\mathcal{C}, \mathcal{Q}\} \rightarrow \mathcal{A}$) also enforces a clean separation between causal reasoning and answering. This prevents answer-level gradients from distorting the reasoning process, preserving the causal integrity of the intermediate representation. This separation is what allows the system to function as a reusable reasoning engine, capable of generalizing to out-of-domain datasets and novel video distributions.

8.3. Positioning Our Work Within the Landscape of Causal AI

While classical causal inference frameworks—such as Structural Causal Models (SCMs)—focus on modeling interventions and counterfactuals, our system operates in the fully observed, post-hoc regime. The causal chain represents the actualized factual path of events within the video. This makes the model distinct from counterfactual reasoning systems, but still firmly within the broader family of causal AI: it explicitly represents, learns, and reasons about causal relations, not mere correlations. In practice, this bridges the gap between causal interpretability and large-scale visual understanding.

8.4. Implications for Future Causal Reasoning Models

Learnable causal reasoning modules exhibit stronger open-world generalization because causal chains provide transferable, event-level causal structure. Our results (especially in Sec 5.5, Table 6, Fig. 6) indicate that large-scale causal-chain supervision can produce models that are not only better at VideoQA but also more generalizable and interpretable. The CCE's strong transfer to out-of-domain datasets suggests the emergence of reusable causal abstractions—patterns of cause–effect reasoning learned from one domain and applied to another. This opens avenues for developing general-purpose causal reasoning engines applicable beyond VideoQA, including robotics, embodied agents, and interactive video systems.

9. Discussion on the Scope of CauCo metric

While our CauCo metric is designed to evaluate the causal coherence of generated causal chains in open-world VideoQA, it is important to clarify the scope and intended interpretation of the term "causal guarantees." In the context of this work, causal guarantees refer specifically to guarantees of causal coherence in the generated reasoning trace, rather than formal guarantees in the sense of interventional or counterfactual causal inference.

CauCo is trained to holistically/comprehensively judge whether a causal chain exhibits logically valid cause–effect progression: whether each event meaningfully leads to the next, whether actors and entities remain consistent, and whether temporal ordering is preserved. This enables CauCo to detect a wide range of causal perturbations—including event removal, order reversal, negation, semantic substitution, and agent swapping. In this way, CauCo provides a robust, open-world mechanism for verifying that a model-generated causal chain maintains internal causal validity, even when evaluated on unseen or diverse video scenarios. Such guarantees are essential for VideoQA systems deployed in unconstrained environments, where causal reasoning traces must remain structurally sound despite domain variability.

However, these guarantees should not be interpreted as formal causal identification guarantees in the framework of Structural Causal Models (SCMs) or do-calculus. CauCo does not verify whether a chain corresponds to the true causal mechanism underlying the video, nor does it reason about interventions or counterfactuals. Instead, it evaluates whether the submitted chain itself adheres to coherent causal logic. Thus, while CauCo strengthens the reliability and interpretability of causal-reasoning modules in open-world VideoQA, it operates at the level of causal consistency, not at the level of causal discovery or causal effect estimation.

By clarifying this distinction, we aim to prevent potential misinterpretation of CauCo's role. Our use of the term "causal guarantees" reflects a practical, representation-level notion of causal soundness in natural-language reasoning chains—one that aligns with the goals of explainable and robust VideoQA—rather than a promise of formal causal inference in the classical statistical sense.

10. CoT and distinction from it

CoT vs. Causal Chains. Causal chains explicitly encode cause–effect relationships, describing how one event leads to another. In contrast, Chain-of-Thought (CoT) is an inference-time reasoning technique in which a model breaks down a complex problem into smaller intermediate steps to improve interpretability and accuracy. Although both terms include the word "chain," they refer to entirely different concepts: causal chains represent causal structure, whereas CoT represents procedural reasoning steps generated during inference.

Model-Level Operationalization vs. Inference-Time Technique. Our approach adopts the CoT philosophy of decomposing complex tasks, but operationalizes it through model-level architectural design rather than inference-time prompting. Specifically, we decompose the overall Causal-Why VideoQA task into two modules: causal reasoning and answer generation. This modularization embodies the CoT principle structurally, distinguishing our method from traditional CoT prompting, which operates solely at inference time.

Accuracy and Video-Groundedness of Reasoning Traces vs. CoT Outputs. Each module is trained via supervised fine-tuning using ground-truth reasoning traces. These traces are accurate, video-grounded, and explicitly capture actions, intentions, and cause–effect relationships. In contrast, CoT outputs generated during inference may be approximate, noisy, or hallucinated, and are not guaranteed to

align with the underlying visual evidence.

11. Further details human study design

Interface design. We have an example of the interface used to record human survey in Fig. 7.

Complementarity across studies. The three user studies are intentionally interdependent: each targets a distinct but related dimension of human interaction with explainable systems—understanding, trust, and preference. By analyzing these dimensions separately, we can disentangle where explanations genuinely add value from where they might only appear persuasive.

Study I: Explainability — understanding over persuasion. In Study I, participants evaluate explanations independent of correctness. They judge whether causal chains help them understand a model's reasoning process, not whether the final prediction is right. This design already tests the cognitive utility of explanations. If participants simply favored "more words," they would have shown a uniform bias toward explanations across all cases—but the results show that explanations were not rated helpful in $\sim$29% of examples. This variability indicates that users are discerning and that preference in Study III cannot be explained by superficial eloquence alone.

Study II: Trustworthiness — calibrating trust, not amplifying it. Study II evaluates trust without ground truth. Here, users decide which system they would rely on, based on how convincing or transparent its reasoning appears. The fact that trust increased in 62% of cases—but decreased when the explanation contradicted participants' expectations—shows that explanations do not automatically inflate trust. Instead, trust is conditional and reflective, suggesting that participants engaged critically with the content rather than reacting to presentation style.

Study III: Preference — interpretability under equal performance. Building on these insights, Study III isolates preference when both systems are equally accurate. Having established that explanations can aid understanding (Study I) and calibrate trust (Study II), this final study explores whether users value such interpretability when correctness is held constant. Together, these studies demonstrate that preference for our explainable system emerges from genuine cognitive and experiential benefits—not from rhetorical or stylistic bias.

Viewed together, the three studies form a coherent narrative: users find explanations helpful (Study I), calibrate their trust appropriately (Study II), and prefer transparency even when performance is equal (Study III). This convergence suggests that our explainable model's advantage arises from meaningful interpretability rather than presentation effects.

Model	NextQA	CVQA	CausalChaos!	Avg.
Baseline 3B	60.23	72.11	67.65	66.66
Baseline 8B	19.41	66.77	08.63	31.60

Table 7. **Naively increasing the model parameters does not help.** We can see that larger model performs inferior to its smaller counterpart, indicating that larger number of parameters alone does not help, but our design and structured approach are crucial in boosting the performance.

Additional Discussion on Study III

In Study III, we restricted evaluation to cases where both the black-box and our explainable model produced the correct answer. This choice was crucial to decouple explanation preference from performance bias.

If one model were more accurate, participants might naturally favor it—not because of its explainability, but because it was right. By holding accuracy constant, we ensured that any differences in human preference stemmed solely from the perceived usefulness, clarity, or satisfaction provided by the explanations.

This design allows us to measure the intrinsic value of explainability—how much users appreciate or trust a system's reasoning process when accuracy is no longer a differentiating factor. Thus, the results of Study III directly reflect the human-centered benefit of transparency rather than performance-driven bias.

Similar methodological controls have been adopted in explainable AI evaluations [11], where isolating the effect of explanations requires holding model accuracy constant. Our setup follows this principle to avoid conflating interpretability with predictive success.

12. Naively increasing parameters does not help

We conducted an experiment demonstrating that naively scaling up model size does not improve performance, highlighting the need for a structured approach with causal chains as effective intermediate representations.

For this, we investigate whether simply increasing model parameters improves performance. To this end, we compare a larger Vision-Language Model (VLM), VILA 1.5 8B, against its smaller counterpart, VILA 1.5 3B—Our model uses this smaller 3B version. For a fair comparison, we fine-tune both models using the same experimental setup. The results, presented in Table 7, indicate that scaling up model size alone does not necessarily enhance performance. This underscores the necessity of a modular approach with causal chains as effective intermediate representations.

Question : why did tom tilt his head?

System A Output :

Prediction : Tom wanted to empty the water inside his ear.

System B Output :

Predictions : Tom wanted to empty the water inside his ear.

Model Reasoning Chain: Tom felt a stream of water hitting his face -> Tom tilted his head to drain the water

Qn1) Which from which system are you able to infer/understand/explain why the answer was chosen?

○ **A** ○ **B** ○ **No preference**

Qn2) Which from which system are more confident in the models choice of answer?

○ **A** ○ **B** ○ **No preference**

Figure 7. **Example of human survey collection interface.**

13. Discussion on computational efficiency

Although our modular framework consists of multiple stages, both training and inference remain computationally efficient. Notably, video data are processed exclusively within the first module, in both phases. The modular training procedure is efficient because only the first stage—responsible for multimodal understanding—requires video processing. This stage produces compact yet semantically rich representations in the form of explicit natural-language causal chains. The subsequent reasoning stage operates entirely on language inputs, obviating the need for multimodal computation.

Consequently, only the first module incurs multimodal processing costs, which are comparable to, or even lower than, those of state-of-the-art systems. For instance, our causal-chain extractor comprises approximately 3 billion parameters, whereas contemporary models such as Qwen2.5VL-7B exceed twice this scale. During inference, the same efficiency advantage holds: only the first module engages in video processing, and it remains significantly smaller and more efficient than competing multimodal architectures.

Moreover, the sequential execution of modules ensures low computational and memory overhead. The framework allows modules to be dynamically loaded and unloaded to accommodate constrained memory environments. For example, the first module can be loaded to extract causal chains, after which the resulting compact representations can be cached and the module released from memory. Subsequently, the second-stage reasoning module can be loaded to perform inference directly over the cached causal chains—without requiring access to the original video data.

Lastly, in the previous section, we further observed that our modular causal chain–guided reasoning approach enables substantially more efficient parameter utilization compared to existing architectures.

14. Extended Details on Base or Source Datasets

Since no existing dataset includes causal chain or reasoning trace annotations to support our approach, we construct causal chains by annotating three challenging causal video QA datasets: 1) **NextQA** [54], 2) **CausalVidQA** [23], and 3) **CausalChaos!** [38]. We call these three datasets as **base datasets** or source datasets. We focus on Causal-Why QA in these datasets.

Preliminary. Base datasets contain triplets of:

1. **Video clip**
2. **Human-written Causal-Why Question on the Video clip mentioned above**
3. **Human-written Gold Correct Answer to the Causal-Why Question and grounded in the Video clip mentioned above**

In the following, we summarize the characteristics of the base datasets individually.

1. **NextQA** is a widely used and challenging VideoQA dataset that includes descriptive, temporal, and causal questions. For our study, we focus solely on causal questions. NextQA features natural videos depicting everyday object interactions without restrictions on specific actors or activities. It sources its videos from the VidOR dataset, selecting 6,000 longer and more interaction-rich videos. These videos primarily capture scenes of family time, children playing, social gatherings, outdoor activities, pets, and musical performances.
2. **CausalVidQA** is a large-scale dataset featuring multiple question types, including Why, How, Who, and What. It is built on the Kinetics dataset, the largest human action

video dataset, which contains videos spanning 700 human actions, including single-person activities, human-object interactions, and human-human interactions. For our dataset, we focus exclusively on causal-Why questions.

3. **CausalChaos!** is a challenging VideoQA dataset designed specifically for causal-Why questions. It is based on Tom & Jerry cartoons, in contrast to the previous two datasets, which feature real-world videos. CausalChaos! presents complex reasoning, frequent scene changes, diverse reasoning types, & varied visual & motion dynamics. These three datasets collectively provide comprehensive scenario coverage.

We follow the original training-validation-test split ratios of each dataset.

15. Further details on evaluation

In this section, we briefly explain the standard evaluation protocol for multi-choice video question answering; and how we ensure fair evaluation protocol consistency.

15.1. Standard Evaluation Protocol for Multiple-Choice Video Question Answering

We evaluate model performance on the Video Question Answering (VideoQA) task using the standard multiple-choice (MC) formulation. Each evaluation instance consists of a video clip $\mathcal{V}$, an associated natural-language question $\mathcal{Q}$, a set of K candidate answers $\{\mathcal{A}_1, ..., \mathcal{A}_K\}$, and a single ground-truth correct answer $\mathcal{A}^* \in \{\mathcal{A}_1, ..., \mathcal{A}_K\}$. The objective of the model is to select the correct answer from the provided candidates. Note that candidate answers and the ground-truth correct answer are provided by the base datasets, not by us.

15.2. Prediction Procedure

Given an input pair $(\mathcal{V}, \mathcal{Q})$, all the models computes a score or probability distribution over all answer candidates. Let s_i denote the compatibility score assigned to candidate $\mathcal{A}_\rangle$. The predicted answer is obtained via:

$$\hat{a} = \arg\max_{a_i} \; s_i \tag{1}$$

15.3. Primary Evaluation Metric

Performance is quantified using answer accuracy, defined as the proportion of instances in which the predicted answer matches the ground-truth answer:

$$\text{Accuracy} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}[\hat{a}_j = a_j^*] \tag{2}$$

where N denotes the number of examples in the evaluation split, and $\mathbf{1}[\cdot]$ is the indicator function.

Accuracy is the standard evaluation metric for MCQA as mentioned in the main paper Section 5.

15.4. Fair Evaluation Protocol Consistency

All models are trained and evaluated using the official dataset splits of the base dataset to ensure comparability with prior work and complete fairness. During evaluation, models must select exclusively from the provided answer candidates, and no external supervision or retrieval is permitted. Furthermore, evaluation uses the same human-authored QA data for all the models. Note that the Causal chains predicted by the CCE module of our model serve only as intermediate reasoning scaffolds in our model, not evaluation targets ensuring complete fairness. Also that during the test time or inference time, the causal chains are predicted by the CCE module of our model—our model does not make use of groundtruth causal chains to ensure complete fairness.

16. Further Implementation Details

For LLaMA-3.1-8b finetuning we make use of the library available on HuggingFace to perform LoRA finetuning. We finetune using a Paged AdamW with a learning rate of 2e-4, with the following LoRA configuration: Rank=16, alpha=32, dropout=0.05, target_modules=['up_proj', 'down_proj', 'gate_proj', 'k_proj', 'q_proj', 'v_proj', 'o_proj'].

We use causal language modeling (CLM) loss as loss objective during the stage-1 training or the Loss-1 in Figure 3 main paper. Causal language modeling loss is widely adopted and is especially suitable for generating causal chains grounded in videos because it enforces a forward-moving, step-by-step prediction process that mirrors how causal reasoning unfolds. By allowing the model to condition only on past video and question tokens, CLM encourages coherent temporal understanding and prevents "peeking" ahead, making each reasoning step depend on the previous one. This aligns naturally with interpreting events in order, linking causes to effects, and producing clear, grounded explanations for video-based questions. We follow VILA third stage finetuning recipe provided in their official repository, but with a lower learning rate of 1e-5. For Stage-2, we use cross entropy loss on the predicted answer choice and the correct groundtruth answer defined in the base datasets.

We follow the procedure provided in [38] to generate semantically modified/perturbed chains for main paper Sections 4.3 and 5.2.

17. Dataset Stats

Samples from each dataset are as follows: 1) NextQA (16874); 2) CausalVidQA (24205); 3) CausalChaos!

(4945).

18. Prompting Details

We design various to be various stages and modules of our approach and experiments. Details and samples of these prompts are provided in the following. They are also included in our code to be released.

In the initial construction of the causal chains or reasoning steps for SFT of CCE (Section 3 main paper), we provide the question-correct answer pairs as inputs to the Oracle LLM (GPT-4o model) with the prompt as seen in Fig 8

During the SFT process, we provide the video, corresponding question via the prompt seen in 9 to the selected CCDA model.

For our one-shot chain generation experiment, we provide a video and the corresponding ground truth chain from the train set as a further demonstration example to guide the models and invoke appropriate knowledge in them as shown in Fig 10.

19. Extended Related Work

In the following, we discuss the widely used state-of-the-art VideoQA models. We have discussed their central concepts and unique design characteristics.

Traditional Models.

• *BlindQA* [2]. In this approach, no visual information is leveraged. Answers are chosen directly based on the questions. In a nutshell, this model learns a mapping from question to answer. Higher performance by method would suggest that the dataset contains questions that are not visually-grounded.

• *EVQA* [2]. This method extends BlindQA baseline by incorporating the visual stream modeled by an LSTM.

• *Spatio-Temporal Reasoning in Visual Question Answering (STVQA)* [19]. This work introduces three novel video QA tasks that demand spatio-temporal reasoning skills to answer questions accurately. In addition, a new TGIF-QA dataset has been created to facilitate research in this field. To address this issue, a dual-LSTM-based approach with both spatial and temporal attention mechanisms has been proposed as a baseline model.

• *Motion-Appearance Co-Memory Networks (CoMem)* [15]. A novel Video QA framework, combining Dynamic Memory Network (DMN) principles with motion and appearance features. This innovative approach leverages a co-memory attention mechanism to incorporate both motion and appearance cues. It employs a temporal conv-deconv network to create multi-level contextual information and utilizes a dynamic fact ensemble method for constructing dynamic temporal representations tailored to specific questions.

• *Heterogeneous Memory Enhanced Multimodal Attention Model (HME)* [12].This innovative end-to-end trainable Video QA framework begins by generating global context-aware visual and textual features. It achieves this by interacting the current inputs with memory contents. Subsequently, it integrates these multimodal features through attentional fusion to make accurate inferences for answering questions.

• *HCRN* [21]. This is a hierarchical framework with conditional relation networks as building blocks models input video at multiple scales (clip-, full video-level) in a cascaded manner. Visual features at each level are conditioned on the question features. The joint representation is fed into the classifier for answer prediction.

• *HGA* [20]. Leverages heterogeneous graph reasoning module and a co-attention unit to capture the local and global correlations between video clips, linguistic concepts and their cross-modal correspondences.

• *Multimodal Iterative Spatial-temporal Transformer (MIST)* [14]. MIST, designed for long-form Video Question Answering (VideoQA), revolutionizes conventional dense spatial-temporal self-attention. It accomplishes this by utilizing two critical modules: segment and region selection, which adaptively pick out frames and image regions tied to the questions. Following this, it processes diverse visual concepts effectively with an attention mechanism. This process occurs iteratively across multiple layers, empowering the model with multi-event reasoning capabilities.

Vision Language Models.

• *Video-LLaMA* [58]. This work extends LLaMA (Large Language Model Meta AI) with video and audio processing capabilities, enabling video-based reasoning, captioning, and multi-modal understanding

• *VideoChat2* [26]. To improve temporal understanding over prior MLLMs, this work adopts 2 main approaches, a 3-stage progressive multi-modal training of Vision-Language Alignment, Vision-Language Connection, and Instruction Tuning and training on diverse instruction-tuning data of 2M samples from 34 distinct sources incorporating both image and video data.

• *ViLA-1.5* [30]. Employs advanced pre-training strategies, including interleaved image-text data and unfreezing large language models (LLMs) during training.

Causal Video Question Answering (Causal-Why Video QA) Causal video question answering (Causal-Why Video QA) presents unique challenges beyond traditional video QA tasks, as it requires models to infer underlying causes rather than merely describing visual content. Existing Causal-Why Video QA models often struggle with causal reasoning, frequently relying on superficial correla-

What is the causal chain in the following question-answer pair? Please return the concise causal chain in the form of event_A -> event_B -> event_C... without any additional text.

Question: {question}

Answer: {answer}

Figure 8. Prompt sample for generating pseudo ground truth from GPT-4o

<video>

From the video and question: {Question}. Generate a causal chain.

Figure 9. Prompt used in fine-tuning of CCDA models

tions, such as action-verb and object-noun co-occurrences, rather than true cause-and-effect relationships. Several benchmarks and models have been proposed to advance video question answering. The TVQA dataset[22] focuses on multi-choice question answering based on television narratives, incorporating both vision and language understanding. VIOLIN [31] introduces video-and-language inference tasks, emphasizing commonsense reasoning. However, these datasets primarily evaluate high-level inference rather than explicit causal reasoning. Data sets such as CausalChaos [38], NextQA [54] and CausalVidQA [24] demand explicit causal reasoning and more tailored towards causal video question answering.

<video> Why did Tom open his mouth? Generate a causal chain.

Causal Chain: Tom's desire to stop Jerry and Nibbles from stealing the key -> Tom wanting to bite the key -> Tom opening his mouth

<video> {Question} Generate a causal chain.

Causal Chain:

Figure 10. One-shot prompt with example and reference video.

ReasVQA [28] uses automatically generated rationales as auxiliary supervision & is monolithic: it predicts answers directly, with rationale generation serving only as training guidance—not required at inference.	**Ours ChainReaction** is modular, using causal chains in 2-stage pipeline where one module generates the chain and another predicts the answer from it. This creates a structural dependency—perturbing the chain changes the answer (Exp. 5.2, Fig. 4)—which ReasVQA lacks.
Video-STaR [59] & Ours ChainReaction differ in their approach to reasoning. Video-STaR prioritizes scalability with a weakly supervised, human-free pipeline, treating rationales as a byproduct of self-training & accepting them if they yield correct labels. Thus, reasoning emerges from label matching rather than direct supervision.	**ChainReaction** prioritizes causal depth via a human-in-the-loop pipeline, treating reasoning as an explicit, modular mechanism. Using supervision to train models to generate & consume structured causal chains ensures reasoning is a learned requirement.
AoTD [45] generates reasoning traces and distills them into a monolithic Video-LLM; the traces serve as training supervision, but answers do not structurally depend on them.	**Ours ChainReaction** is a two-stage modular pipeline where a causal-chain extractor produces an explicit natural-language causal chain, and the answerer is explicitly conditioned on that chain to produce the answer.
Both focus on reasoning traces: **VideoCoT** [50] uses heuristic scoring & refinement of general rationales, while ours ChainReaction applies explicit causal-chain coherence checks via cross-model verification & human grounding.	
CAGE [42] is an inspiring early effort in the direction of two-stage framework, but is limited to language-only QA with commonsense explanation. Whereas ChainReaction is a video causal-why pipeline that learns an explicit causal event chain (grounded in video) as a supervised intermediate and answers from that structured chain (with stage-wise training).	
ChainReaction extends **Q-ViD** [33]-style two-step models by: 1) using a learned causal event chain as the intermediate representation instead of question-guided frame descriptions; and 2) training dedicated chain extractor, answerer modules on its new causal-chain dataset, whereas Q-ViD is largely prompted/zero-shot with frozen components.	
CauCo follows an **LLM-as-verifier paradigm** (*e.g.*, [56]) but differs in that the verifier is supervised fine-tuned to detect causal incoherence using systematically perturbed negative chains (*e.g.*, actor swaps, negation, order reversal), making the judge explicitly sensitive to cause–effect structure rather than generic textual quality. **To evaluate trustworthiness of CauCo**, we assessed CauCo's alignment with human judgment. Three annotators judged 100 extracted chains for causality. **Human–Human agreement** averaged $\sim$**85%**, and **CauCo–Human agreement** was $\sim$**80%**. Despite causal ambiguity, this near-human agreement supports CauCo as a reliable metric.	