ReaSon: Reinforced Causal Search with Information Bottleneck for Video Understanding

Yuan Zhou[1*], **Litao Hua**[1], **Shilong Jin**[1], **Wentao Huang**[1], **Haoran Duan**[2]
[1]Nanjing University of Information Science and Technology
[2]Tsinghua University
{zhouyuan, 202412621441, 202312490769, 202412491490}@nuist.edu.cn, haoran.duan@ieee.org

Abstract

Keyframe selection has become essential for video understanding with vision-language models (VLMs) due to limited input tokens and the temporal sparsity of relevant information across video frames. Video understanding often relies on effective keyframes that are not only informative but also causally decisive. To this end, we propose ***Re**inforced **Ca**usal **S**earch with Inf**o**rmation Bottlen**e**ck (ReaSon)*, a framework that formulates keyframe selection as an optimization problem with the help of a novel Causal Information Bottleneck (CIB), which explicitly defines keyframes as those satisfying both predictive sufficiency and causal necessity. Specifically, ReaSon employs a learnable policy network to select keyframes from a visually relevant pool of candidate frames to capture predictive sufficiency, and then assesses causal necessity via counterfactual interventions. Finally, a composite reward aligned with the CIB principle is designed to guide the selection policy through reinforcement learning. Extensive experiments on NExT-QA, EgoSchema, and Video-MME demonstrate that *ReaSon* consistently outperforms existing state-of-the-art methods under limited-frame settings, validating its effectiveness and generalization ability.

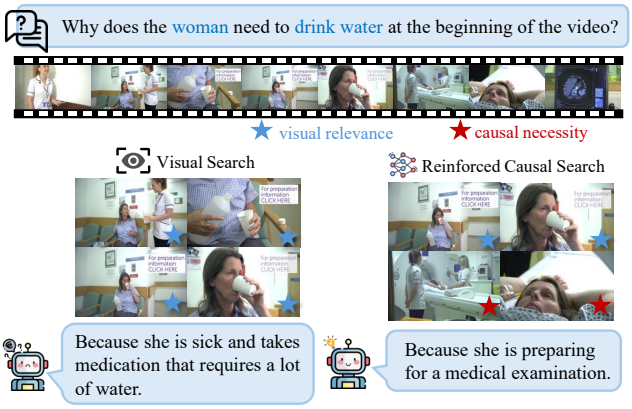

Figure 1: Illustration of limitations of visual relevance and the importance of causal necessity in keyframe selection. The visual search method selects visually relevant frames (blue stars) but misses causally decisive frames (red stars). In contrast, reinforced causal search captures causally necessary frames, leading to more accurate answers.

Introduction

Recent advances in video understanding have been greatly driven by the rise of vision-language models (VLMs) (Tang et al. 2025; Nguyen et al. 2024). However, these models are severely constrained by input token budgets and suffer from the intrinsic redundancy of videos, where informative evidence is often sparsely distributed (Wang et al. 2024; Ye et al. 2025; Cao et al. 2025). As a result, recent studies have increasingly focused on developing frame selection strategies that extract a subset of keyframes (Ye et al. 2025; Wang et al. 2025; Ma et al. 2025; Fan, Guo, and Yang 2025) to improve both computational efficiency and reasoning accuracy for video understanding. Yet a fundamental question remains unresolved: *What defines a keyframe that is essential for video understanding?*

Most current methods (Ye et al. 2025; Guo et al. 2025; Wang et al. 2024) define keyframes as an informative and compact subset of frames. These approaches typically select frames that are visually or semantically aligned with the question or answer, treating such correlation as a proxy for informativeness while restricting the number of frames to achieve compactness (Guo et al. 2025; Ye et al. 2025; Wang et al. 2024). This process implicitly adheres to the Information Bottleneck (IB) principle (Tishby, Pereira, and Bialek 2000), which aims to preserve task-relevant information while discarding redundancy under a compression constraint. However, a high visual or semantic correlation does not guarantee decisive evidence for VLMs reasoning, which results from a lack of causal dependencies. As illustrated in Fig. 1, frames that appear visually relevant may not always be necessary for the correct reasoning process, whereas causally decisive frames, such as prior causes and subsequent effects, are often overlooked.

Motivated by this limitation, we revisit the concept of keyframes from a causal perspective (Yu et al. 2025). Keyframes should meet two essential criteria: 1) predictive sufficiency, which ensures the selected subset supports accurate inference, and 2) causal necessity, which means

*Corresponding author.

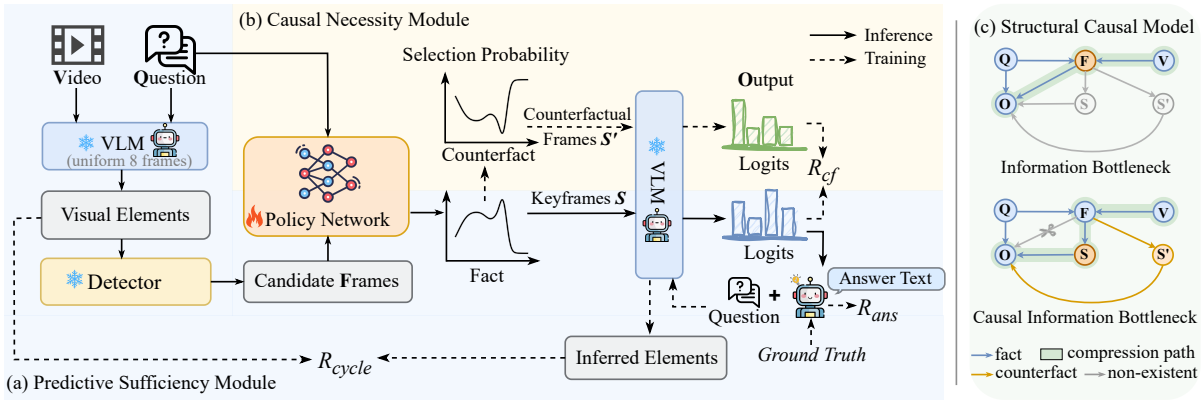

Figure 2: Framework of proposed ReaSon. (a) and (b) illustrate the predictive sufficiency and causal necessity modules, respectively, where a policy network learns to select keyframes based on CIB-aligned rewards. (c) shows the structural causal models. Q and V denote the question and the video, respectively. F and S represent selected frame subsets. S' is a counterfactual selection to assess causal necessity. O means the output. Bottleneck variables are highlighted with orange circles.

no frame in the subset can be removed without impairing the output. To this end, we introduce the Causal Information Bottleneck (CIB) for keyframe selection, which extends the classic IB framework by incorporating an interventional term to capture predictive sufficiency and causal necessity. Furthermore, we present ReaSon, a reinforced causal search method that utilizes a learnable policy network to guide the keyframe selection process grounded in the CIB principle.

ReaSon consists of two core components: a predictive sufficiency module and a causal necessity module. The predictive sufficiency module first constructs a candidate pool by detecting question-relevant visual elements across the video. A learnable policy network is then leveraged to select a compact subset from the candidate pool. This two-stage process allows the model to initially localize potentially relevant regions via visual grounding and subsequently distill a subset of frames that are sufficient for reasoning, thus effectively reducing redundancy. In addition, the causal necessity module evaluates whether each selected frame is causally indispensable. Counterfactual interventions are constructed by altering the selected frame subset to generate a counterfactual input. The resulting distributional changes in the output of the VLM are measured to assess causal necessity, based on the assumption that substantial output shifts imply causal dependence. These measurements provide learning signals that guide the policy network to distinguish frames that merely correlate with the answer from those that are causally decisive. To implement the CIB-guided frame selection, we employ reinforcement learning to train a selection policy and design three rewards aligned with predictive sufficiency and causal necessity.

Finally, we conduct extensive experiments to validate the effectiveness of ReaSon across diverse video understanding scenarios. ReaSon is evaluated on three representative benchmarks: NExT-QA(Xiao et al. 2021) for causal, temporal, and descriptive reasoning; EgoSchema (Mangalam, Akshulakov, and Malik 2023) for egocentric video understanding; and Video-MME (Fu et al. 2025) for long-form video understanding. Our method is compared against recent non-selection-based methods and state-of-the-art frame selection approaches. Results show that ReaSon achieves the best accuracy of 81.4% on NExT-QA validation set and 72.2% on EgoSchema subset with only 8 frames. On Video-MME, ReaSon outperforms state-of-the-art methods, reaching improvements of 2.6% (8 frames) and 2.3% (32 frames). Ablation studies demonstrate that both the sufficiency and necessity modules make substantial contributions to performance. ReaSon consistently improves reasoning accuracy across different base models, highlighting its strong generalization ability.

In summary, our main contributions are as follows:

- We formally formulate keyframe selection as an information-theoretic optimization problem under the CIB, integrating predictive sufficiency and causal necessity.
- We present ReaSon, a CIB-based causal search framework consisting of two dedicated modules that collaboratively identify sufficient and necessary keyframes.
- Three CIB-aligned rewards are introduced to guide the selection policy via reinforcement learning, capturing answer correctness, semantic consistency, and causal necessity.
- ReaSon outperforms existing frame selection and non-selection baselines under limited-frame settings, demonstrating strong effectiveness and generalization across diverse video understanding scenarios.

Method

Preliminary

Formally, let V[1] and Q denote the input video and question as random variables and let F denote a subset of frames selected from the video V. The goal of keyframe selection is to

[1]All uppercase letters denote random variables, and lowercase letters denote their instances unless stated otherwise.

identify F that preserves the relevant information in V and Q for the output O. As illustrated in Fig. 2(c), this process can be described by a structural causal model (SCM) (Pearl 2009). F serves as an intermediate variable that compresses the input V, which is structurally aligned with the Information Bottleneck (IB) principle (Tishby, Pereira, and Bialek 2000). Therefore, the objective of keyframe selection can be naturally formulated under the IB framework as:

$$\max \ \mathcal{I}(F;O) \quad \text{s.t. } \mathcal{I}(V,Q;F) \leq \beta, \tag{1}$$

where β controls the allowed amount of information retained from the input and $\mathcal{I}$ denotes mutual information. This objective encourages the selected frames to be informative for the output and as compressed as possible with respect to the input.

In practice, most existing frame selection methods follow IB principle by selecting a compact subset of frames that preserve high visual relevance to the question and answer. These approaches approximate the objective of maximizing $\mathcal{I}(F;O)$ by prioritizing visually aligned frames under a fixed input budget. For example, VideoAgent (Wang et al. 2024) predicts intermediate event descriptions using a large language model and retrieves frames with high image-text similarity to those events, while T* (Ye et al. 2025) employs object detectors to locate frames containing entities mentioned in the question. However, by focusing solely on visual or semantic correlations, these methods may overlook decisive frames that are indispensable for correct reasoning.

Causal Information Bottleneck

To support reliable reasoning in video understanding, effective keyframes should satisfy two essential criteria:

- **Predictive Sufficiency**: The selected frames must provide enough information to accurately answer the question, yielding outputs consistent with those derived from the full video;
- **Causal Necessity**: The frames should be a minimal subset with no redundancy, which means removing any frame in the subset would significantly affect the output.

To capture both sufficiency and causal necessity, we extend the Information Bottleneck to Causal Information Bottleneck with a causal perspective. The original bottleneck variable F lacks the capacity to represent selection as an intervenable decision, making it unsuitable for analyzing causal necessity. As illustrated in Fig. 2(c), a new variable S, which denotes the target keyframes, is introduced to replace F as the information bottleneck. This structural adjustment isolates the causal effect of selection and enables formal analysis of both predictive sufficiency and necessity. The Causal Information Bottleneck objective can be defined as follows:

$$\max \ \mathcal{I}(S;O) + \mathcal{I}_c(O;\mathrm{do}(S)) \quad \text{s.t. } \mathcal{I}(V,Q;S) \leq \beta, \tag{2}$$

where $\mathrm{do}(S)$ denotes an intervention on S, and $\mathcal{I}_c$ represents mutual information defined under causal interventions. $\mathcal{I}(S;O)$ encourages predictive sufficiency, ensuring the selected keyframes retain enough information to infer the correct output. The second term $\mathcal{I}_c(O;\mathrm{do}(S))$ quantifies the influence of keyframes S under interventions to measure the causal necessity. The constraint $\mathcal{I}(V,Q;S) \leq \beta$ limits the information capacity, preventing overly redundant selections.

Reinforced Causal Search

Building on the CIB, we propose **ReaSon**, a reinforced causal search approach that employs a learnable policy $\pi_\theta(S \mid F, Q)$ to select keyframes S. As illustrated in Fig. 2, ReaSon comprises two dedicated modules targeting predictive sufficiency and causal necessity, respectively. However, mutual information terms in the CIB are intractable to compute or differentiate in practice. Therefore, in this section, tractable approximations of the CIB objective are derived in detail and are connected to three rewards. These distinct rewards aligned with the CIB objective jointly guide policy learning via reinforcement learning.

Predictive Sufficiency Module This module focuses on optimizing the first term of the CIB objective, $\mathcal{I}(S;O)$. $\mathcal{I}(S;O)$ can be expressed as:

$$\mathcal{I}(S;O) = \mathbb{E}_{p(s,o)}\left[\log \frac{p(o \mid s)}{p(o)}\right]. \tag{3}$$

Directly computing Eq. (3) is intractable in general cases. To derive a tractable surrogate, we introduce a variational distribution $q_\phi(o \mid s)$ to approximate the true posterior $p(o \mid s)$ (Alemi et al. 2017; Kingma, Welling et al. 2013), and rewrite Eq. (3) as:

$$\begin{aligned}
\mathcal{I}(S;O) &= \mathbb{E}_{p(s,o)}\left[\log \frac{q_\phi(o \mid s)}{p(o)} + \log \frac{p(o \mid s)}{q_\phi(o \mid s)}\right] \\
&= \mathbb{E}_{p(s,o)}\left[\log \frac{q_\phi(o \mid s)}{p(o)}\right] \\
&+ \ \mathbb{E}_{p(s)}\left[D_{\mathrm{KL}}(p(o \mid s) \,\|\, q_\phi(o \mid s))\right] \\
&\geq \mathbb{E}_{p(s,o)}\left[\log \frac{q_\phi(o \mid s)}{p(o)}\right].
\end{aligned} \tag{4}$$

Furthermore, ignoring the marginal term $p(o)$, which is independent of s, yields a looser but tractable approximation:

$$\mathcal{I}(S;O) \gtrsim \mathbb{E}_{p(s,o)}[\log q_\phi(o \mid s)] \triangleq J_1(s). \tag{5}$$

The resulting surrogate objective $J_1(s)$ provides a tractable approximation of predictive sufficiency, which serves as a foundation for the reward design in practice. As shown in Fig. 2(a), given a specific video v and a question q, we first construct a candidate frame pool $f = \{f_1, f_2, \ldots, f_M\} \subseteq v$ using a heuristic-based pre-selection strategy inspired by T* (Ye et al. 2025), which filters out visually irrelevant frames. M denotes the size of the candidate frame pool. Uniformly sampled frames are fed into a frozen VLM to extract target visual elements E_q. The visual elements are then matched against all frames using an open-vocabulary detector to form the candidate pool. Then, a policy network $\pi_\theta(S \mid F = f, Q = q)$ is introduced to model the distribution over possible frame subsets conditioned on the candidate pool f and the question q. The policy network

assigns a selection probability to each frame in the candidate pool f, and a keyframe subset $s \sim \pi_\theta(S \mid f, q)$ is then sampled via a multinomial process, subject to a cardinality constraint $|s| \leq K$. K denotes the maximum number of selected frames. This acts as a practical proxy to enforce the constraint $\mathcal{I}(V, Q; S) \leq \beta$. The selected subset s with the question q is passed into the frozen VLM, which serves as an implementation of the variational distribution $q_\phi(o \mid s)$, to generate an answer. A binary reward is computed by comparing the generated answer with the ground truth:

$$R_{\text{ans}} = \mathbb{I}[\text{VLM}(s, q) = \text{gt}], \tag{6}$$

where $\text{VLM}(s, q)$ denotes the generated textual output and gt is the ground truth. This reward serves as an approximation of the mutual information objective $\mathcal{I}(S; O)$, encouraging the policy to select frames that lead to a correct answer.

To further reinforce predictive sufficiency, we introduce a cycle consistency reward that encourages semantic alignment throughout the reasoning process. After producing the final answer from the selected keyframes, the predicted answer is concatenated with the original question and sent back to the VLM to infer a set of visual elements, denoted as E_a. Notably, the video frames are not accessible during this stage. The answer-based elements E_a are compared with the previously extracted target elements E_q to assess whether the reasoning process completes a semantic cycle: from visual input to answer reasoning and back to visual attribution. The cycle consistency reward is defined as:

$$R_{\text{cycle}} = \text{IoU}(E_q, E_a). \tag{7}$$

A strong alignment between E_q and E_a indicates that the selected keyframes successfully preserve the semantic cues required to answer the question. This reward complements the answer reward R_{ans}, providing additional guidance from the perspective of semantic consistency.

Causal Necessity Module This module is designed to optimize the second term $\mathcal{I}_c(O; \text{do}(S))$ in the CIB, which can be expanded as follows:

$$\mathcal{I}_c(O; \text{do}(S)) = \mathbb{E}_{s \sim \pi_\theta} \left[D_{\text{KL}} \left(p(o \mid \text{do}(s)) \,\|\, p(o) \right) \right]. \tag{8}$$

According to the Causal Markov condition (Pearl 2009), once the direct causes of a variable are fixed, the variable is conditionally independent of all other variables that are not its effects or direct causes. In our case, the output O is directly determined by the selected frames S. Under the assumption that the structural mechanism from S to O remains unchanged under intervention (Pearl 2009), we can treat $do(S = s)$ as equivalent to conditioning on $S = s$, and rewrite Eq. (8) as:

$$\mathcal{I}_c(O; \text{do}(S)) = \mathbb{E}_{s \sim \pi_\theta} \left[D_{\text{KL}} \left(p(o \mid s) \,\|\, p(o) \right) \right]. \tag{9}$$

To address the intractability of computing the marginal distribution $p(o)$ and assess the causal necessity of selected frames, a counterfactual selection strategy is introduced to approximate $p(o)$ by inverting the original selection policy, denoted as $\tilde{\pi}$. The counterfactual strategy can be defined as:

$$\tilde{\pi}(f_i) = \frac{1 - \pi_\theta(f_i)}{\sum_{j=1}^{M} \left(1 - \pi_\theta(f_j)\right)}, \; i, j \in [1, \ldots, M], \tag{10}$$

where $\pi_\theta(f_i)$ and $\tilde{\pi}(f_i)$ denote the original and the counterfactual selection probability assigned to the i-th frame, respectively. As shown in Fig. 2(b), a counterfactual subset s' is sampled according to $\tilde{\pi}$, which serves as a contrastive sample. Therefore, $p(o)$ is computed as:

$$p(o) = \mathbb{E}_{s' \sim \tilde{\pi}} \left[p(o \mid s') \right]. \tag{11}$$

Finally, substitute this approximation into Eq. (9) and obtain the following objective:

$$\mathcal{I}_c(O; \text{do}(S)) = \mathbb{E}_{s \sim \pi_\theta} \left[D_{\text{KL}} \left(p(o \mid s) \,\|\, \mathbb{E}_{s' \sim \tilde{\pi}}[p(o \mid s')] \right) \right]. \tag{12}$$

To reduce the computational cost during reinforcement learning, we approximate the expectation over counterfactual selections using a single Monte Carlo sample (Alemi et al. 2017) and rewrite Eq. (12) as:

$$\begin{aligned} \mathcal{I}_c(O; \text{do}(S)) &\approx \mathbb{E}_{s \sim \pi_\theta,\, s' \sim \tilde{\pi}} \left[D_{\text{KL}} (p(o \mid s) \,\|\, p(o \mid s')) \right] \\ &\triangleq J_2(s, s'). \end{aligned} \tag{13}$$

In practice, to make the optimization of the surrogate objective $J_2(s, s')$ feasible, a counterfactual reward is defined to guide the selection policy. Specifically, o and o' denote the logits outputs from the VLM given inputs (s, q) and (s', q), respectively. The reward can be formulated as:

$$R_{\text{cf}} = D_{\text{KL}} \left(\text{softmax}(o) \,\|\, \text{softmax}(o') \right). \tag{14}$$

This reward measures the effect of counterfactual interventions on the output to capture the causal necessity. Greater divergence implies that the absence of keyframes leads to the significant changes.

Optimization

The CIB objective can be finally approximated as:

$$\max_{s \sim \pi_\theta} J(s, s') = J_1(s) + J_2(s, s') \quad \text{s.t. } |s| \leq K. \tag{15}$$

We construct a composite reward function R aligned with $J(s, s')$ as a practical proxy that enables implicit optimization of the CIB objective via reinforcement learning. The reward R is defined as a weighted combination of the three rewards introduced above:

$$R = R_{\text{ans}} + \lambda_1 R_{\text{cycle}} + \lambda_2 R_{\text{cf}}. \tag{16}$$

To train the policy π_θ, we employ a group-wise policy gradient method (Chu et al. 2025), which estimates gradients based on multiple sampled selections per training instance. Specifically, for each video and question, we sample G subsets of keyframes $\{s_i\}_{i=1}^{G} \sim \pi_\theta$ by multinomial sampling, along with a counterfactual subset $s' \sim \tilde{\pi}$ for comparison. The corresponding model outputs o_i and o' are obtained via VLM and the rewards R_i are computed as defined in Eq. (16). To stabilize learning, intra-group advantages $\hat{A}_i$ are calculated by mean-centering rewards, as defined in Eq. (17), which reduces gradient variance while avoiding the bias and instability introduced by standard deviation normalization.

$$\hat{A}_i = R_i - \frac{1}{G} \sum_{j=1}^{G} R_j. \tag{17}$$

Finally, the selection policy is updated via policy gradient:

$$\nabla_\theta \mathcal{L} = \frac{1}{G} \sum_{i=1}^{G} \hat{A}_i \cdot \nabla_\theta \log \pi_\theta(s_i \mid f, q). \tag{18}$$

Method	VLM	Mean Frames	NExT-QA				EgoSchema
			Tem	Cau	Des	Avg	
Non-selection Methods							
MVU (Ranasinghe et al. 2025)	Mistral-13B	16	55.4	48.1	64.1	55.2	60.3
LangRepo (Kahatapitiya et al. 2024)	Mistral-8×7B	180	51.4	64.4	69.1	60.9	66.2
VideoChat2 (Li et al. 2024a)	GPT-4	16	57.4	61.9	69.9	61.7	54.4
LLoVi (Zhang et al. 2024a)	GPT-4	–	61.0	69.5	75.6	67.7	61.2
VideoINSTA (Liao et al. 2024)	GPT-4	90	–	–	–	72.3	65.0
Frame Selection Methods							
VideoAgent (Wang et al. 2024)	GPT-4	8.4	64.5	72.7	81.1	71.3	60.2
VideoAgent (Fan et al. 2024)	GPT-4	–	60.0	76.0	76.5	70.8	62.8
LVNet (Park et al. 2024)	GPT-4o	12	65.5	75.0	81.5	72.9	68.2
DrVideo (Ma et al. 2025)	GPT-4	0.5fps	–	–	–	–	66.4
VideoTree (Wang et al. 2025)	GPT-4	63.2	70.6	76.5	83.9	75.6	66.2
AKEYS (Fan, Guo, and Yang 2025)	GPT-4o	26.7	72.9	79.0	86.1	78.1	68.6
T* (Ye et al. 2025)	LLaVA-OneVision-7B	8	–	–	–	<u>80.4</u>	66.6
Ours							
ReaSon	Qwen2.5-VL-7B	8	<u>76.4</u>	<u>81.0</u>	<u>86.6</u>	<u>80.4</u>	68.0
ReaSon	LLaVA-Video-7B	8	**77.3**	**82.1**	**87.4**	**81.4**	<u>69.0</u>
ReaSon	GPT-4o	8	70.6	80.2	83.6	77.6	**72.2**

Table 1: Comparison of ReaSon with existing state-of-the-art methods on NExT-QA and EgoSchema. We adopt accuracy (%) as the metric. Results of baseline methods are directly cited from their respective publications. The best result is highlighted in bold, and the second-best is marked with underline.

Experiments

Experimental Setup

Datasets We train ReaSon on NExT-QA (Xiao et al. 2021) training set and evaluate its performance on NExT-QA validation set, EgoSchema subset (Mangalam, Akshulakov, and Malik 2023) and Video-MME (Fu et al. 2025). NExT-QA consists of 5,440 videos, which is designed to test temporal and causal reasoning over short videos. EgoSchema contains 5000 egocentric three-minute videos paired with multiple-choice questions, but only provides public labels for a subset of 500 questions. Video-MME is a recent-proposed long video understanding dataset, with an average video duration of 44 minutes. These datasets cover different video types and reasoning styles, including causal, temporal, egocentric and long video understanding.

Implementation Details All videos are sampled at 1 fps in our proposed method. We leverage YOLO-World (Cheng et al. 2024) as the detector to match visual elements in predictive sufficiency module. BLIP (Li et al. 2022) is used to encode both video frames and questions as the input to the policy network. The policy network consists of a three-layer LSTM (Hochreiter and Schmidhuber 1997) and an MLP (Rumelhart, Hinton, and Williams 1986). During training, we set the candidate frame pool size to 32 and select 8 keyframes ($M = 32$ and $K = 8$), using LLaVA-Video-7B (Zhang et al. 2024d) as VLM. λ_1 and λ_2 are set to 0.5. The number of groups G is set to 4. For inference, we employ LLaVA-Video-7B, Qwen2.5-VL-7B (Bai et al. 2025), and GPT-4o (Hurst et al. 2024) across all datasets. For NExT-QA and EgoSchema, the keyframe selection settings remain consistent with those used during training. For Video-MME, the candidate frame pool size is increased to 64, with 32 keyframes selected to accommodate long video understanding ($M = 64$ and $K = 32$). All 8-frame experiments are conducted on an RTX 3090 GPU, while 32-frame inference is performed on an A100 GPU.

Comparison with Existing Approaches

Table 1 shows a comparison of existing state-of-the-art methods and ReaSon on NExT-QA. We compare our method with both non-selection and frame selection approaches. ReaSon achieves the highest overall accuracy on NExT-QA (81.4% with LLaVA-Video-7B) and state-of-the-art performance on EgoSchema (72.2% with GPT-4o). Compared to non-selection methods such as VideoINSTA using 90 frames, ReaSon outperforms VideoINSTA by 9.1% and 7.2% under 8 frames, highlighting the importance of keyframe selection. Among frame selection methods, ReaSon also achieves top performance with fewer or comparable frames. Compared to the previous SOTA method AKEYS, ReaSon improves accuracy by 3.3% on NExT-QA and 3.6% on EgoSchema. Notably, the largest gains are observed on temporal and causal questions, with an improvement of 4.4% and 3.1% over AKEYS, showing the advantage of our method in modeling causal necessity. Furthermore, under the same frame setting, ReaSon outperforms T* by a substantial 5.6% on EgoSchema. The best performance on each dataset is obtained using different VLMs equipped with our method, which is expected

Method	VLM	Frames	Short	Medium	Long	Overall
Video-LLaVA (Lin et al. 2023)	Vicuna-7B v1.5	8	45.3	38.0	36.2	39.9
LongVA (Zhang et al. 2024b)	Qwen2-7B	8	55.1	46.3	42.1	47.9
GPT-4o	GPT-4o	8	55.7	54.3	51.4	53.8
T* (Ye et al. 2025)	GPT-4o	8	56.4	**57.3**	**56.4**	56.5
ReaSon	LLaVA-Video-7B	8	63.7	48.7	47.4	53.3
Reason	GPT-4o	8	**65.9**	57.1	54.4	**59.1**
LongVA	Qwen2-7B	32	61.1	48.8	45.4	51.8
LLaVA-NeXT-Video (Zhang et al. 2024c)	LLaVA-NeXT-Video-34B	32	61.7	50.1	44.3	52.0
GPT-4o	GPT-4o	32	68.3	60.7	56.3	61.8
T*	GPT-4o	32	69.5	63.5	**59.3**	64.1
ReaSon	LLaVA-Video-7B	32	69.2	55.0	49.3	57.9
ReaSon	GPT-4o	32	**76.8**	**64.2**	58.2	**66.4**
Video-XL (Shu et al. 2025)	Qwen2-7B	128	64.0	53.2	49.2	55.5
VideoChat-Flash (Li et al. 2024b)	Qwen2-7B	512	–	–	55.4	65.3
VideoLLaMA 3 (Zhang et al. 2025)	Qwen2.5-7B	180	80.1	63.7	54.9	66.2
Gemini 1.5 Pro	Gemini 1.5 Pro	1/0.5 fps	81.7	74.3	67.4	75.0
GPT-4o	GPT-4o	384	80.0	70.5	65.3	71.9
Qwen2-VL	Qwen2-VL-72B	768	80.1	71.3	62.2	71.2
LLaVA-Video	LLaVA-Video-72B	64	81.4	68.9	61.5	70.6

Table 2: Comparison of different methods on Video-MME without subtitles. We report accuracy (%) across three video duration categories: Short (<2 minutes), Medium (4-15 minutes), and Long (30-60 minutes). All baseline results are reported as cited from their respective publications. Our method are highlighted with a gray background. Best results in each setting are shown in bold. Noting that methods displayed in gray utilize significantly more frames and proprietary large models that are not publicly available or reproducible, making direct comparisons challenging.

Method	NExT-QA				EgoSchema
	Tem	Cau	Des	Avg	
w/ R_{ans}	76.3	81.1	84.4	80.1	66.0
w/ R_{ans} + R_{cycle}	76.8	81.7	84.3	80.5	68.2
w/ R_{ans} + R_{cycle} + R_{cf}	77.3	82.1	87.4	81.4	69.0

Table 3: Ablation study evaluating the contribution of different reward components in **ReaSon** on NExT-QA and EgoSchema. Results are reported using **LLaVA-Video-7B** with 8-frame input.

VLM	NExT-QA				EgoSchema
	Tem	Cau	Des	Avg	
LLaVA-Video-7B	76.3	81.4	83.9	80.2	65.2
+*ReaSon*	77.3	82.1	87.4	81.4	69.0
Qwen2.5-VL-7B	75.9	80.6	86.0	79.9	65.8
+*ReaSon*	76.4	81.0	86.6	80.4	68.0
GPT-4o	64.4	75.2	76.6	72.0	70.0
+*ReaSon*	70.6	80.2	83.6	77.6	72.2

Table 4: Evaluation of the effectiveness and generalization ability of **ReaSon** across different VLMs on NExT-QA and EgoSchema. **ReaSon** is conducted under 8-frame input.

given the variation in data distribution and the robustness of VLMs. Importantly, ReaSon achieves strong and stable performance across all VLMs, consistently ranking among the top-performing methods on both datasets.

Additionally, we evaluate ReaSon's performance on Video-MME for long video understanding. As shown in Table 2, ReaSon with GPT-4o achieves the best overall accuracy of 59.1% and 66.4% among all methods under comparable configurations with 8 and 32 keyframes, respectively. This highlights the effectiveness of our selection strategy under limited frame budgets. For short videos, ReaSon demonstrates the most significant improvement. Under the 8-frame setting, ReaSon boosts GPT-4o performance from 55.7% to 65.9% and outperforms T* by 9.5%. With 32 frames, ReaSon exceeds GPT-4o and T* by 8.5% and 7.3%, respectively. For medium and long videos, ReaSon exhibits steady improvements. Using 8 frames, it improves GPT-4o performance by 2.8% and 3.0% on medium and long videos, respectively. When extended to 32 frames, the gains over vanilla GPT-4o reach 3.5% for medium and 1.9% for long videos. ReaSon surpasses T* by 0.7% on medium videos but falls short on long videos. Despite certain limitations on long videos, ReaSon remains competitive across diverse video lengths, even compared to methods with significantly larger frame budgets such as Video-XL, VideoChat-Flash, and VideoLLaMA 3.

Figure 3: The visualization of frame selection results demonstrates the effectiveness of our approach compared to the previous state-of-the-art method *T**. Our approach pays less attention to irrelevant regions (in gray) and identifies more causal decisive keyframes.

Ablation Study

We conduct an ablation study to assess the effectiveness of reward components in ReaSon. As shown in Table 3, each reward component in ReaSon contributes to the overall performance. The cycle consistency reward R_{cycle} improves sufficiency by aligning visual semantics, while the counterfactual reward R_{cf} enhances necessity by encouraging the selection of frames that are causally necessary for reasoning. The complete reward combination achieves the best results on both datasets.

We further evaluate the effectiveness and generalization of ReaSon with different VLMs. Table 4 shows the results using LLaVA-Video-7B, Qwen2.5-VL-7B, and GPT-4o. Comparing each VLM with and without ReaSon, we observe consistent performance improvements across all settings, confirming the effectiveness of ReaSon as a plug-in module for enhancing video understanding. The largest overall gains are observed on Qwen2.5-VL-7B (+3.8% on EgoSchema) and GPT-4o (+5.6% on NExT-QA). While different VLMs show dataset-specific strengths and no single model performs best across all scenarios, ReaSon does not rely on any specific model and can enhance each VLM, highlighting its robustness and generalization.

Visualization

Fig. 3 presents the visualization of frame selection results on a video sampled from EgoSchema, paired with a manually crafted question. Compared to the previous state-of-the-art method T*, our approach identifies more relevant and causally decisive frames and correctly answers the question. During the selection process, our method anchors regions that are highly relevant to the question and also attends to their preceding and succeeding context to capture potential causal cues. As a result, the overall frame selection probability exhibits three prominent peaks. While T* is also able to localize question-relevant regions, its frame selection tends to include unnecessary frames, leading to broader and less precise peaks. In contrast, our method pays less attention to irrelevant regions and maintains only minimal focus on unrelated regions to avoid missing useful information. Within the highly relevant regions, ReaSon effectively captures causal necessity and filters out frames that are correlated but not essential.

Conclusion

In this paper, we introduce ReaSon, a reinforced causal search framework for video understanding, grounded in a novel Causal Information Bottleneck. By modeling keyframe selection as an optimization of both predictive sufficiency and causal necessity, ReaSon captures not only visually relevant but also causally decisive information. Through a dual-module architecture and reinforcement learning guided by a composite reward, our method identifies compact keyframe subsets that support accurate reasoning. Extensive experiments across diverse video types and question categories demonstrate that ReaSon consistently outperforms strong baselines under limited-frame settings while maintaining strong generalization across different VLMs. We believe our framework offers a principled and extensible foundation for efficient video understanding. In future work, we plan to further improve frame selection for more challenging long video scenarios.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Cao, S.; Zhang, Z.; Jiao, J.; Qiao, J.; Song, G.; Shen, R.; and Meng, X. 2025. MASR: Self-Reflective Reasoning through Multimodal Hierarchical Attention Focusing for Agent-based Video Understanding. arXiv:2504.17213.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16901–16911.

Chu, X.; Huang, H.; Zhang, X.; Wei, F.; and Wang, Y. 2025. GPG: A Simple and Strong Reinforcement Learning Baseline for Model Reasoning. arXiv:2504.02546.

Fan, S.; Guo, M.; and Yang, S. 2025. Agentic Keyframe Search for Video Question Answering. arXiv:2503.16032.

Fan, Y.; Ma, X.; Wu, R.; Du, Y.; Li, J.; Gao, Z.; and Li, Q. 2024. VideoAgent: A Memory-Augmented Multimodal Agent for Video Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 75–92.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24108–24118.

Guo, W.; Chen, Z.; Wang, S.; He, J.; Xu, Y.; Ye, J.; Sun, Y.; and Xiong, H. 2025. Logic-in-Frames: Dynamic Keyframe Search via Visual Semantic-Logical Verification for Long Video Understanding. arXiv:2503.13139.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Kahatapitiya, K.; Ranasinghe, K.; Park, J.; and Ryoo, M. S. 2024. Language Repository for Long Video Understanding. arXiv:2403.14622.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes. arXiv:1312.6114.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 12888–12900.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; Wang, L.; and Qiao, Y. 2024a. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22195–22206.

Li, X.; Wang, Y.; Yu, J.; Zeng, X.; Zhu, Y.; Huang, H.; Gao, J.; Li, K.; He, Y.; Wang, C.; Qiao, Y.; Wang, Y.; and Wang, L. 2024b. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling. arXiv:2501.00574.

Liao, R.; Erler, M.; Wang, H.; Zhai, G.; Zhang, G.; Ma, Y.; and Tresp, V. 2024. VideoINSTA: Zero-Shot Long Video Understanding via Informative Spatial-Temporal Reasoning with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP*, 6577–6602.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. arXiv:2311.10122.

Ma, Z.; Gou, C.; Shi, H.; Sun, B.; Li, S.; Rezatofighi, H.; and Cai, J. 2025. DrVideo: Document Retrieval Based Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18936–18946.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-Form Video Language Understanding. *Advances in Neural Information Processing Systems*, 36: 46212–46244.

Nguyen, T.; Bin, Y.; Xiao, J.; Qu, L.; Li, Y.; Wu, J. Z.; Nguyen, C.-D.; Ng, S.-K.; and Tuan, L. A. 2024. Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives. arXiv:2406.05615.

Park, J.; Ranasinghe, K.; Kahatapitiya, K.; Ryu, W.; Kim, D.; and Ryoo, M. S. 2024. Too Many Frames, Not All Useful: Efficient Strategies for Long-Form Video QA. arXiv:2406.09396.

Pearl, J. 2009. *Causality*. Cambridge university press.

Ranasinghe, K.; Li, X.; Kahatapitiya, K.; and Ryoo, M. 2025. Understanding Long Videos in One Multimodal Language Model Pass. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. Learning Representations by Back-Propagating Errors. *Nature*, 323(6088): 533–536.

Shu, Y.; Liu, Z.; Zhang, P.; Qin, M.; Zhou, J.; Liang, Z.; Huang, T.; and Zhao, B. 2025. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26160–26169.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video Understanding with Large Language Models: A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The Information Bottleneck Method. arXiv:physics/0004057.

Wang, X.; Zhang, Y.; Zohar, O.; and Yeung-Levy, S. 2024. VideoAgent: Long-Form Video Understanding with Large Language Model as Agent. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 58–76.

Wang, Z.; Yu, S.; Stengel-Eskin, E.; Yoon, J.; Cheng, F.; Bertasius, G.; and Bansal, M. 2025. VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3272–3283.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T. 2021. Next-QA: Next Phase of Question-Answering to Explaining Temporal Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9777–9786.

Ye, J.; Wang, Z.; Sun, H.; Chandrasegaran, K.; Durante, Z.; Eyzaguirre, C.; Bisk, Y.; Niebles, J. C.; Adeli, E.; Fei-Fei, L.; Wu, J.; and Li, M. 2025. Re-thinking Temporal Search for Long-Form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8579–8591.

Yu, X.; Wang, Z.; Yang, L.; Li, H.; Liu, A.; Xue, X.; Wang, J.; and Yang, M. 2025. Causal Sufficiency and Necessity Improves Chain-of-Thought Reasoning. arXiv:2506.09853.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; Jin, P.; Zhang, W.; Wang, F.; Bing, L.; and Zhao, D. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. arXiv:2501.13106.

Zhang, C.; Lu, T.; Islam, M. M.; Wang, Z.; Yu, S.; Bansal, M.; and Bertasius, G. 2024a. A Simple LLM Framework for Long-Range Video Question-Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 21715–21737.

Zhang, P.; Zhang, K.; Li, B.; Zeng, G.; Yang, J.; Zhang, Y.; Wang, Z.; Tan, H.; Li, C.; and Liu, Z. 2024b. Long Context Transfer from Language to Vision. arXiv:2406.16852.

Zhang, Y.; Li, B.; Liu, H.; Lee, Y. J.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024c. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/. Accessed: 2024-07-30.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024d. Video Instruction Tuning with Synthetic Data. arXiv:2410.02713.