The VLM as Sensor: Bayesian Active Search for Long Video Understanding

Chong Tang
University of Bristol
University of Southampton
Bristol, United Kingdom
chong.tang@bristol.ac.uk

Sannara Ek*
University of Cambridge
Cambridge, United Kingdom
se521@cam.ac.uk

Dirk Koch
University of Manchester
Manchester, United Kingdom
dirk.koch@manchester.ac.uk

Robert Mullins
University of Cambridge
Cambridge, United Kingdom
robert.mullins@cl.cam.ac.uk

Alex Weddell
University of Southampton
Southampton, United Kingdom
asw@ecs.soton.ac.uk

Jagmohan Chauhan*
University College London
UCL AI Centre
London, United Kingdom
jagmohan.chauhan@ucl.ac.uk

Abstract

What if the best way to search a video with a vision-language model (VLM) is to stop asking the model to search? Recent approaches give the VLM control over temporal navigation, tying search quality to the model's reasoning ability. We propose the opposite: BeliefSearch treats the VLM as a *noisy sensor* and delegates navigation to an external Bayesian controller. A single belief over video segments anchors the system: the question sets the prior, each VLM observation updates it, and the same belief decides whether to search, which segment to examine next, and when to stop. Because the belief is explicit, we can trace why each choice was made. The same signal also drives training: each turn earns credit for the uncertainty it reduces, and the final reward is gated by how much the search lowered entropy. This blocks a reward-hacking shortcut where the model otherwise learns to skip search and guess from the preview frames. On long-video benchmarks, our method outperforms all prior search-based methods by up to 8.8 points while using 5 to 13$\times$ fewer frames than recent RL search baselines.

1 Introduction

Long video understanding is, at its core, a search problem. Given a question about a video that may span minutes to hours [Fu et al., 2025, Wu et al., 2024, Zhou et al., 2025], a system must find the relevant temporal evidence and decide when it has enough to answer. The video is too long to process in full within current VLM context limits [Bai et al., 2025, Chen et al., 2024], so some form of temporal navigation is unavoidable.

A growing line of work addresses this by giving the model itself control over where to look. Some train it to navigate via reinforcement learning (RL) [Pan et al., 2025]. Others equip it with temporal

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

grounding tools [Ye et al., 2025] or orchestrate multi-step analysis through agent pipelines [Wang et al., 2024, 2025, Chen et al., 2025]. In each case, the VLM must simultaneously perceive visual content *and* decide where to look. But these are different skills: VLMs are trained to describe what they see [Zhai et al., 2023, Li et al., 2024, Zhang et al., 2024c], while navigation is an additional capability added through extra RL training [Pan et al., 2025]. When the two are coupled, search quality becomes bound to model capability: a weaker model navigates worse, and errors compound across turns. What is missing is an approach that lets VLMs focus on perception while an external system handles the search.

Our Bayesian active search framework, BeliefSearch, **treats the VLM as a noisy sensor and delegates searching to an external controller**. The entire system is built around a single object: a belief distribution over video segments. The question initializes it as a prior over where the answer might lie. Each VLM description, a short natural-language report on a segment, reshapes the belief as new evidence. The belief's properties then determine every downstream decision: which segment to examine next, when to stop, and whether search is needed at all. Because this basis is explicit rather than hidden in model weights, a user can see what evidence the controller collected and why each segment was examined.

The same belief state also provides a natural training signal. Standard approaches like GRPO [Shao et al., 2024, Guo et al., 2025] assign one episode-level reward across all turns, leaving models unable to tell which search steps helped. Process reward models [Lightman et al., 2023] address per-step credit in language reasoning but require training a separate verifier. Our belief-entropy reduction supplies a per-turn signal (information gain) at no extra cost, and the same entropy gates the reward to remove a reward-hacking failure where models would otherwise learn to skip search. Our contributions are:

- A Bayesian active search framework that decouples perception from navigation: the VLM describes what it sees while a single belief drives routing, selection, and stopping, yielding interpretable search traces and using 5 to $13\times$ fewer frames than recent RL search baselines.
- An observation model that pairs NLI-scored likelihoods with asymmetric belief propagation, encoding video's evidence structure: positive evidence is local, negative is global.
- Temporal-Difference (TD) Belief Learning: belief entropy serves as a free critic for per-turn credit, and an evidence-grounded reward removes the reward-hacking failures of standard multi-turn RL.

2 Related Work

Static frame selection for long video. Single-pass frame selection picks frames once before any QA step. Approaches range from uniform sampling [Fu et al., 2025], question-conditioned selection with SigLIP [Zhai et al., 2023], and adaptive keyframe sampling [Tang et al., 2025] to diversity-based DPP selection [Kulesza and Taskar, 2012] and learned localizers [Yu et al., 2023]. Memory compression [Song et al., 2024, He et al., 2024] and extended context windows [Liu et al., 2024, Zhang et al., 2024b] reduce what must be discarded. These approaches scale well, but if the selected frames miss the critical evidence, no later step can recover it.

Model-controlled temporal search. Iterative methods address this by viewing the video in turns: the model examines a segment, produces a response, and selects the next. TimeSearch-R [Pan et al., 2025] trains this navigation via reinforcement learning. Video-R1 [Feng et al., 2025] is an alternative chain-of-thought RL method for video reasoning. T* [Ye et al., 2025] equips GPT-4o with temporal grounding tools, while VideoAgent [Wang et al., 2024], VideoTree [Wang et al., 2025], LVAgent [Chen et al., 2025], and caption-then-reason pipelines like LLoVi [Zhang et al., 2024a] orchestrate multi-step analysis through LLM pipelines. These iterative methods focus on relevant regions, but in every case the model controls the search, so segment choices stay hidden in internal state and the trace is hard to audit.

Separating the sensor from the controller. In other domains, separating perception from control is foundational: the controller chooses what to observe, the sensor reports what it sees. The idea has roots in Bayesian experimental design and information-gathering POMDPs [Chaloner and Verdinelli, 1995, Kaelbling et al., 1998], where agents maintain beliefs and select actions to reduce uncertainty. Modern instances span embodied agents [Brohan et al., 2023, Huang et al., 2023] and Bayesian

experimental design with language models [Choudhury et al., 2025]. Long video raises three conceptual challenges that classical Bayesian active search does not directly address: the latent state is defined by a natural-language question rather than a structured world model, observations arrive as free-form language rather than calibrated measurements, and evidence is correlated across time rather than independent across turns. §4 develops the machinery for a VLM sensor and a Bayesian controller, with each component answering one of these challenges.

3 Problem Setting

Long videos hold sparse question-relevant evidence: the answer to a given question often depends on a few specific moments. We formalize the task as Bayesian active search over temporally coherent chunks $c_1, \ldots, c_N$, where the system targets moments most likely to contain the answer. A belief distribution $\mathbf{b} = (b_1, \ldots, b_N)$ assigns each chunk the probability that it holds the evidence for the answer. The belief evolves by Bayes' rule: each observation enters as a likelihood that converts the prior into a posterior. At every turn, the system picks a chunk to observe, integrates the result, and decides whether to stop. The belief is the only state carrying information between turns.

Chunks. We search at chunk granularity rather than frame granularity because events span many frames but usually stay within a single scene, so scene-aligned segments correspond roughly to atomic units of evidence. This keeps the belief over a manageable number of semantically meaningful units rather than a long list of redundant frames. SigLIP [Zhai et al., 2023] encodes each frame, and we group frames into chunks by detecting scene boundaries from drops in consecutive-frame cosine similarity. Each chunk c_i carries an L2-normalized embedding $\mathbf{E}_i \in \mathbb{R}^d$ aggregated from its frames, and the question is encoded the same way as $\mathbf{e}_q \in \mathbb{R}^d$. These embeddings provide an initial signal about how well the question aligns with each chunk.

Search Pipeline. The search loop opens with a preview turn: the VLM sees a few frames from every chunk and rephrases the question as a declarative hypothesis h (see Appendix B for the full design rationale). At each subsequent turn, the controller selects a chunk c_i, prompts the VLM with sampled frames from c_i, and receives a natural-language description o_i of what it saw. A scorer compares o_i against h and returns a numeric signal, which becomes the likelihood that updates the belief. The loop continues until a stopping criterion is reached or the turn budget is exhausted. Figure 1 traces the full path.

This formulation decomposes the search loop into a set of explicit design choices. The update needs a prior, a likelihood, and a rule for evidence propagation across chunks. The controller needs a way to decide whether the loop is needed, a selection policy, and a stopping rule. Training has to make the VLM's descriptions useful to the scorer. §4 develops each in turn.

4 Method

4.1 The Bayesian System

Prior. The cosine similarities $s_i = \mathbf{E}_i^\top \mathbf{e}_q$ decompose orthogonally into a mean and a spread. The mean across the N chunks, $\bar{s} = \frac{1}{N}\sum_i s_i$, measures overall question-video fit, and the spread measures how sharply the question discriminates between chunks. We separate the two: the spread shapes the prior, while $\bar{s}$ feeds the routing decision in §4.2. Subtracting the mean and rescaling by $\sqrt{d}$ matches the $1/\sqrt{d}$ noise scale of random cosines in d dimensions, placing the logits on a question-independent scale so a single τ works across questions. We then softmax into the prior:

$$\ell_i = \frac{(s_i - \bar{s})\sqrt{d}}{\tau}, \qquad b_i^{(0)} = \frac{\exp(\ell_i)}{\sum_j \exp(\ell_j)}. \tag{1}$$

Here ℓ_i is the per-chunk logit, $\tau > 0$ a fixed temperature, and $b_i^{(0)}$ the initial belief.

When the prior concentrates on a few chunks, the rest contribute little to the search. The perplexity $N_{\text{eff}} = 2^{H(\mathbf{b}^{(0)})}$, with $H(\mathbf{b}) = -\sum_i b_i \log_2 b_i$, gives the effective number of chunks the prior covers. We merge adjacent low-mass chunks until N reaches $\max(3, \lfloor\sqrt{N_{\text{eff}}} + \frac{1}{2}\rfloor)$: $\sqrt{N_{\text{eff}}}$ shrinks the bin count sub-linearly, and 3 is the empirical bound we keep.

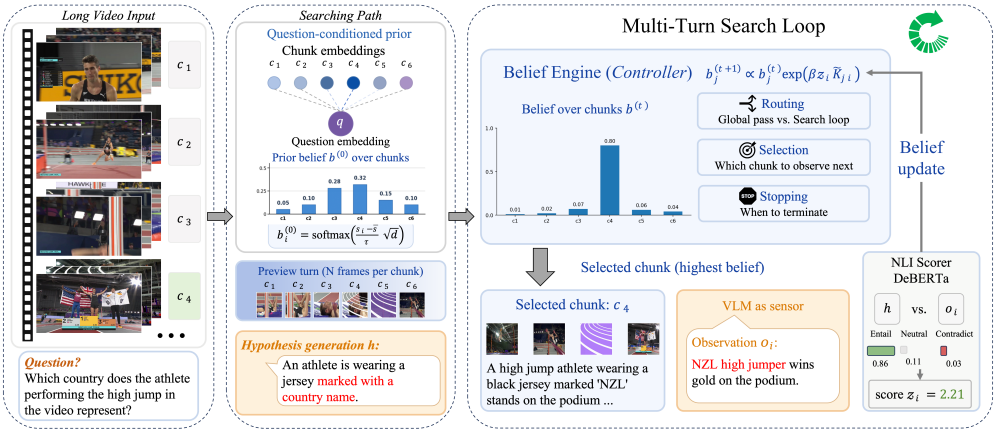


Figure 1: BeliefSearch pipeline. A belief over chunks drives routing, selection, and stopping. At each turn, the VLM describes the selected chunk, and an NLI scorer compares the description against the hypothesis h to produce the likelihood that updates the belief.

Likelihood. Each VLM description o_i enters the belief as a multiplicative factor, which requires a scorer to turn o_i into a number and a functional form for the factor itself. Start with the scorer. Letting the VLM score its own observation (as agentic video search does [Pan et al., 2025, Feng et al., 2025, Wang et al., 2024, 2025, Ye et al., 2025]) risks a feedback loop in which the VLM's miscalibration compounds across turns. A cheaper alternative scores by question-chunk similarity (e.g., SigLIP cosine in AKS [Tang et al., 2025]) at no extra cost, but similarity is direction-blind: "the man picked up the cup" and "the man did not pick up the cup" score almost identically. We use an NLI cross-encoder (DeBERTa-v3-large [He et al., Bowman et al., 2015, Williams et al., 2018]) instead, following the precedent of NLI-based fact verification [Thorne et al., 2018]. Trained on millions of entailment pairs, it scores each observation o_i against the hypothesis h from the preview turn and returns a three-way distribution over entailment, neutral, and contradiction. The three classes give direction, and temperature scaling (Appendix A) reduces NLI's overconfidence. We reduce the three-way distribution to a log-odds score:

$$\text{score}(o_i;\, h) = \log \frac{p_{\text{entail}}(o_i, h)}{1 - p_{\text{entail}}(o_i, h)}. \tag{2}$$

Let $z_i = \text{score}(o_i;\, h)$ denote the score at chunk c_i, a signed value in log-likelihood units that the multiplicative belief update consumes. The form of the multiplicative factor as a function of z is derived from three properties: identity at $z = 0$, monotonicity in z, and (since search runs over multiple turns) multiplicativity under independent observations (two scores z_1 and z_2 producing the same belief change as a single $z_1 + z_2$). By Cauchy's functional equation, the unique form satisfying these is

$$\exp(\beta z), \tag{3}$$

with $\beta > 0$ setting evidence strength.

Belief Update. Applying the likelihood only to the selected chunk treats positive and negative z_i identically, but the two make different claims. A positive z_i is an *existence* claim (c_i contains the answer), which does not transfer to neighbors: finding the answer at one location says nothing about whether it also lies at another. A negative z_i is an *absence* claim (c_i does not), which does propagate: the probability mass we remove has to land somewhere, and where it lands depends on which other chunks resemble c_i. The Bayesian update therefore takes a sign-dependent form:

$$b_j^{(t+1)} \propto b_j^{(t)} \cdot \exp(\beta z_i \cdot \tilde{K}_{ji}), \qquad \tilde{K}_{ji} = \begin{cases} \mathbb{1}[j = i] & z_i > 0 \\ K_{ji} & z_i \le 0, \end{cases} \tag{4}$$

where the propagation kernel K_{ji} describes how mass shed from c_i should redistribute across chunks. We use a blend of temporal proximity (events near c_i in time inherit more) and semantic similarity (chunks with embeddings close to $\mathbf{E}_i$ inherit more):

$$K_{ji} = \max\left(0,\ \alpha \cdot \exp\left(-\frac{|m_j - m_i|}{\sigma}\right) + (1 - \alpha) \cdot \mathbf{E}_j^\top \mathbf{E}_i\right), \tag{5}$$

with $\alpha \in [0, 1]$, m_j the temporal midpoint of c_j, and σ a bandwidth set by per-chunk duration. Figure 5 (Appendix A) illustrates the asymmetry.

4.2 Three Decisions from One Belief

With the belief defined and the update rule in place, we start building the controller. Three decisions follow at three moments of the search: *routing* before the first VLM call, *selection* at each search turn, and *stopping* when further turns are no longer informative.

Routing. Do we always need to run the search loop? Some questions can be answered without it, either because the video is short enough that one VLM call sees all of it, or because the question asks about the video as a whole rather than a specific moment. The routing decision has to be made before any VLM call, so the only signals available are those already computed during belief setup. We combine two of them into a routing score:

$$\text{route_global} = \left(\underbrace{\min\left(1, \frac{N_0}{N_{\text{raw}}}\right)}_{\text{coverage}} + 2 \cdot \bar{s} \right) > \theta, \tag{6}$$

where N_{raw} is the pre-merge scene count and N_0 is the chunk count at which a single global pass densely covers the video. Coverage measures budget adequacy: when $N_{\text{raw}} \leq N_0$, a single VLM call sees the video at adequate resolution and search has nothing to add. The other term is the mean similarity $\bar{s}$ from §4.1. Because both are by-products of belief setup, routing adds no extra computation. When the score clears the threshold, the controller takes a single global pass over frames selected by a DPP.

Selection. At each search turn, the controller picks the highest-belief chunk among those not yet observed. The belief is a probability-weighted ranking of where the answer is likely to lie, updated with every observation, so its argmax is the natural target.

Stopping. A fixed turn budget does not adapt to how the belief evolves: it uses too many calls when the belief has settled quickly and too few when the belief is still moving. The belief itself gives a cleaner signal. At the start of every turn, we check two completion conditions, and either one ends the search. *Entropy convergence*: the normalized entropy $H(\mathbf{b}^{(t)})/\log_2 N$ drops below a threshold, meaning probability mass has concentrated on a single chunk or a small group and further observations are unlikely to change the argmax. *Noise patience*: several consecutive observations return no positive evidence, meaning the scorer has exhausted what it can find on this question and additional turns would add nothing.

4.3 TD-Belief Learning

The same entropy that decides when to stop can also tell us how to train. Standard post-training methods like GRPO [Shao et al., 2024] assign a single episode-level reward and normalize by the group mean, a bandit baseline suited to single-turn RLHF. But our rollouts can span up to 8 turns, so every turn receives the same advantage, and progress cannot be attributed to specific turns. PPO [Schulman et al., 2017] restores per-turn credit through a learned value function, but fitting that critic on sparse episode-level rewards is harder, especially across variable-length trajectories (Appendix D summarizes the differences).

The belief itself can play the role of a value function. The controller maintains a belief state $\mathbf{b}^{(t)}$ at every turn $t \in \{0, 1, \ldots, T\}$, where T is the terminal turn at which the model emits an answer. Its entropy measures uncertainty about where the answer lies, and our reward rises as that uncertainty falls (§4.3). Setting $V^{(t)} = -H(\mathbf{b}^{(t)})$ gives an analytic critic that aligns with what a learned value function would track, with no network to train. With the standard convention $V^{(T)} = 0$ at termination, the per-turn temporal-difference (TD) error is

$$\delta_t = H(\mathbf{b}^{(t)}) - H(\mathbf{b}^{(t+1)}), \tag{7}$$

the information gain at turn t. Turns that reduce uncertainty earn positive advantage, and turns that do not are penalized. At the terminal turn, the reward enters through $\delta_{T-1} = R_{\text{norm}} - V^{(T-1)} =$

$R_{\text{norm}} + H(\mathbf{b}^{(T-1)})$, where R_{norm} is the GRPO group-normalized terminal reward. Generalized Advantage Estimation [Schulman et al., 2015] then produces per-turn advantages,

$$A_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l}, \tag{8}$$

with $\gamma = 1, \lambda = 0.95$. The advantage at turn t applies to all tokens in the VLM's response at that turn, giving dense per-turn credit on top of GRPO's group-relative normalization.

An evidence-grounded reward. Per-turn credit solves only half the problem. The other half is the reward shape, and two conventional choices we tried failed in different ways. (i) Accuracy only: episode-level accuracy does not depend on evidence, so grounded answers and lucky guesses receive the same credit. The model kept searching but learned no preference for evidence-driven trajectories. (ii) Accuracy plus an efficiency bonus: the efficiency term now actively rewarded fewer turns, and the model responded by skipping search and answering from the preview frames. We gate accuracy credit by the fraction of entropy the search reduces:

$$R = w_f \cdot R_{\text{format}} + w_a \cdot R_{\text{accuracy}} \cdot \underbrace{\left(1 - \frac{H(\mathbf{b}^{(T)})}{H(\mathbf{b}^{(0)})}\right)}_{\text{belief alignment}}. \tag{9}$$

A correct answer that fails to reduce entropy, such as one obtained by preview guessing, earns no credit.

5 Experiments

5.1 Setup

Training data and procedure. Training has two stages: SFT on 26,322 distilled traces, then GRPO under TD-Belief GAE. Roughly three-quarters of the traces come from **LongVideoHaystack** [Ye et al., 2025] (9,378 traces, 30–60-minute ego videos, multiple-choice) and **LongViTU** [Wu et al., 2025] (10,815 traces, 10–60-minute videos, open-ended). The remainder comes from **LLaVA-Video-178K** [Zhang et al., 2024c], **ActivityNetQA** [Yu et al., 2019], and **CinePile** [Rawal et al., 2024], which add the shorter and varied content the global pass sees at inference. The teacher is Qwen3-VL-235B-A22B, running BeliefSearch in the inference-time configuration. Training uses up to 256 NVIDIA GH200 GPUs (96 GB each), and inference runs on a single GH200. Filter criteria and hyperparameters are in Appendix A.

Benchmarks and baselines. The evaluation is conducted on three benchmarks: **VideoMME** (VME) [Fu et al., 2025], **LongVideoBench** (LVB) [Wu et al., 2024], and **MLVU** [Zhou et al., 2025] (multiple-choice subset, n=2,174), spanning balanced multi-genre QA, long-context multi-event reasoning, and heterogeneous task categories. The baselines fall into 4 categories. *Search and agentic methods* (TimeSearch-R, Video-R1, T*, VideoAgent, VideoTree) are the direct model-as-navigator competitors. *Single-pass baselines* (uniform, SigLIP top-k, DPP selection) test whether iterative search adds anything beyond a strong single-pass budget of 256 frames. *Standard VLMs* (Qwen2.5-VL-7B, InternVL2.5-8B, LLaVA-Video-7B) establish what comparable open-source VLMs achieve in a single pass over uniformly sampled frames. *Proprietary models* (GPT-4o [Hurst et al., 2024], Gemini-1.5-Pro [Team et al., 2024]), evaluated the same way, serve as a ceiling reference.

Models. The primary VLM is Qwen3-VL-8B. We also report Qwen2.5-VL-7B [Bai et al., 2025], the base used by TimeSearch-R and Video-R1, to control for the base VLM. SigLIP-SO400M [Zhai et al., 2023] encodes frames and questions, and DeBERTa-v3-large [He et al.] scores observations. Both auxiliary models are frozen.

5.2 Results

Main results. Table 1 reports accuracy across all 3 benchmarks. BeliefSearch on Qwen3-VL-8B beats the strongest search-based baselines (TimeSearch-R, Video-R1) by 5 to 9 points on VME and LVB while using far fewer frames (Table 4), and matches or exceeds the single-pass baselines at 256 frames. On VME-S, single-pass baselines lead by about 1 point: 256 frames already cover a short

Table 1: Main results. Best results in **bold**. VME = VideoMME (with short/medium/long splits VME-S/M/L), LVB = LongVideoBench. †: reported by the method's paper. ‡: secondary citation.

Method	Base	VME-S	VME-M	VME-L	**VME**	**LVB**	**MLVU**
Our method							
BeliefSearch	Qwen3-VL-8B	79.8	69.6	67.4	**72.3**	**66.9**	**72.0**
BeliefSearch	Qwen2.5-VL-7B	75.0	67.5	58.2	66.9	61.5	71.7
Search / agentic methods							
TimeSearch-R	Qwen2.5-VL-7B	76.8†	67.1†	56.0†	66.6†	60.1†	71.5†
Video-R1	Qwen2.5-VL-7B	74.1†	65.1†	55.6†	66.9†	58.1†	68.4†
T*	GPT-4o	69.5†	63.5†	59.3†	64.1†	—	—
VideoTree	GPT-4	67.8†	59.9†	54.2†	—	—	60.4†
VideoAgent	GPT-4	—	—	49.0‡	56.0‡	—	—
Single-pass baselines							
Uniform 256f	Qwen3-VL-8B	80.8	68.5	61.8	70.4	67.0	71.7
SigLIP top-k 256f	Qwen3-VL-8B	80.2	67.9	61.2	69.8	66.7	70.3
DPP 256f	Qwen3-VL-8B	80.5	67.6	62.4	70.2	66.3	72.7
Standard VLMs (single pass)							
Qwen2.5-VL-7B	—	76.3†	66.0†	54.6†	65.1†	56.0†	70.2†
InternVL2.5-8B	—	—	—	—	64.2†	60.0†	68.9†
LLaVA-Video-7B	—	—	—	—	58.6‡	58.2‡	70.8‡
Proprietary (single pass)							
GPT-4o	—	—	—	—	77.2†	66.7†	64.6†
Gemini-1.5-Pro	—	—	—	—	75.0†	64.0†	—

Table 2: Component ablations.

Variant	VME (Δ)	LVB (Δ)	MLVU (Δ)
BeliefSearch	**72.3**	**66.9**	**72.0**
Selection policy			
Random chunk selection	66.8 (−5.5)	62.6 (−4.3)	64.3 (−7.7)
Sequential scan	68.3 (−4.0)	60.9 (−6.0)	66.1 (−5.9)
Scoring function			
SigLIP scorer (replace NLI)	67.5 (−4.8)	63.3 (−3.6)	65.3 (−6.7)
Belief propagation			
No kernel	67.1 (−5.2)	63.3 (−3.6)	65.6 (−6.4)
Symmetric kernel	63.3 (−9.0)	58.9 (−8.0)	62.7 (−9.3)
Positive-only kernel	61.3 (−11.0)	54.4 (−12.5)	60.9 (−11.1)
Prior initialization			
Uniform prior (no QC)	65.8 (−6.5)	60.6 (−6.3)	61.3 (−10.7)
Stopping criterion			
Fixed budget ($T_{\max}=8$)	69.4 (−2.9)	66.0 (−0.9)	70.3 (−1.7)

video densely. To control for the underlying VLM, we also train BeliefSearch on Qwen2.5-VL-7B, the same base TimeSearch-R and Video-R1 use: the headline accuracies are nearly tied (66.9 vs 66.6 on VME), but BeliefSearch's advantage concentrates on long content, the regime where active search matters most (+2.2 on VME-L, +1.4 on LVB).

Component ablations. Table 2 removes each component one at a time. The propagation kernel matters most: a symmetric kernel costs 8 to 9 points across benchmarks and a positive-only kernel 11 to 12, while removing the kernel altogether costs only 4 to 6. A wrongly shaped kernel is worse than no kernel: the asymmetry between existence and absence claims (§4.1) is what carries the gain. A uniform prior loses 6 to 11 points, with the strongest effect on MLVU where heterogeneous task types put the most pressure on question conditioning. Replacing NLI with SigLIP drops 4 to 7 points, in line with the direction-blind critique in §4.1. Random or sequential selection drops 4 to 8 points, and the gain over random holds on every one of MLVU's 7 task types (Appendix E). Entropy stopping moves accuracy by 1 to 3 points.

Table 3: Routing analysis. %g = fraction routed to global.

Configuration	VME-S	VME-M	VME-L	**VME**	**LVB**	**MLVU**	%g (V/L/M)	Turns
Search only (no routing)	58.7	67.3	62.6	62.9	65.8	68.5	0/0/0	3.5/3.0/3.9
+Routing [**ours**]	79.8	69.6	67.4	**72.3**	**66.9**	**72.0**	48/42/13	2.2/2.7/3.3
Δ (search→ours)	+21.1	+2.3	+4.8	**+9.4**	+1.1	+3.5		

Table 4: Compute efficiency. Frames/q = frames fed into the VLM per question.

	VME		LVB		MLVU	
Method	Frames/q	Acc	Frames/q	Acc	Frames/q	Acc
Uniform 256f	256	70.4	256	67.0	256	71.7
SigLIP top-k 256f	256	69.8	256	66.7	256	70.3
DPP 256f	256	70.2	256	66.3	256	72.7
BeliefSearch	**91**	**72.3**	**80**	**66.9**	**61**	**72.0**
TimeSearch-R	526	66.6	607	60.1	750	71.5
Video-R1	768	66.9	768	58.1	768	68.4

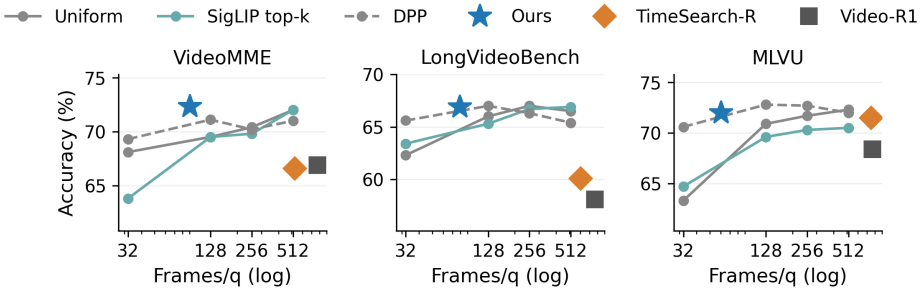


Figure 2: Accuracy–compute Pareto frontier across 3 benchmarks.

Routing. Routing recovers 21 points on VME's short split (Table 3), where the search loop was costing accuracy and not just compute, and 9 points on the full benchmark. On LVB, 42% clear the threshold and accuracy moves 1 point: search and a single global pass are comparable on routed traffic. On MLVU, only 13% clear (most videos too long) and routing still adds 3 points.

Qualitative trajectory. Figure 3 shows belief-guided search recovering from a wrong tentative commitment: BeliefSearch first answers "car" across two early turns showing a car interior, then revises to "helicopter" (the correct answer) at turn 5 after seeing one on a landing pad. Because the controller, not the model, decides when to stop, the early commit does not end the loop. Model-driven search lacks this recovery: TimeSearch-R and Video-R1 each commit to "car" and never revisit. Appendix F groups nine more examples by theme: answer flipping, both baselines fail, methodical elimination, and our own failure modes.

Compute efficiency. BeliefSearch uses 3 to 4 times fewer frames per question than the single-pass baselines and 5 to 13 times fewer than TimeSearch-R and Video-R1 (Table 4), while matching or exceeding their accuracy. BeliefSearch's preview shows two frames per chunk and each turn loads limited frames from one chunk; TimeSearch-R instead re-samples the full video every turn, and Video-R1 takes a single 768-frame pass. Figure 2 shows we Pareto-dominate every baseline across all 3 benchmarks, and Figure 4 confirms the gain is not a global-average effect: on VME our accuracy holds across video durations while Uniform 128f drops on long videos and TimeSearch-R drops at the extremes.

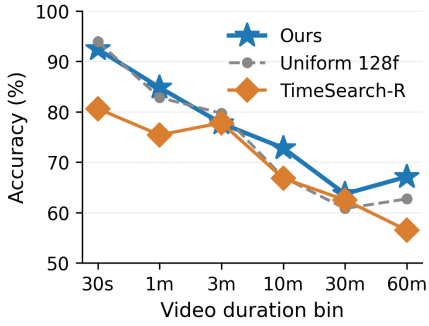


Figure 4: Accuracy vs. video duration on VME.

Training dynamics. Zero-shot on the base model drops more than 30 points because the model has never seen

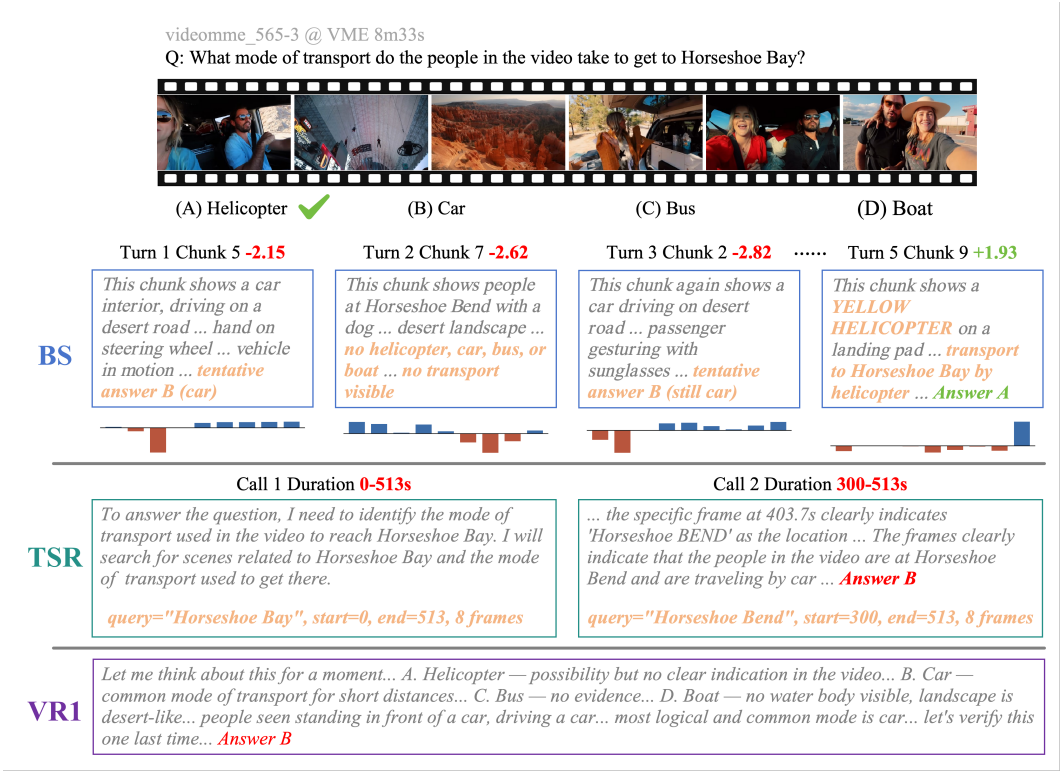


Figure 3: Qualitative results for a representative example. Additional examples are in Appendix F.

Table 5: Training analysis: reward shapes and credit-assignment schemes (reduced training budget).

Training recipe	VME	LVB	MLVU	Avg turns
Naive reward shapes				
Accuracy only	39.0	37.3	42.8	3.2
Accuracy + efficiency bonus	36.5	34.9	40.0	1.5
Evidence-grounded reward				
Episode-level GRPO	56.7	53.9	62.0	2.8
TD-Belief GAE (**ours**)	**59.0**	**56.4**	**64.7**	3.4

the multi-turn protocol. SFT recovers the format but leaves descriptions too thin for the controller to act on, and only the RL stage crosses back over the base model's non-search accuracy (Appendix C). The controller is identical across stages, so the final gain reflects improvement in description quality, not navigation.

Using accuracy alone as the reward, the model performs worse than the non-search baseline (Table 5). Adding an efficiency bonus makes it worse: the search loop collapses to 1.5 turns on average and accuracy drops by another 2 to 3 points, the preview-guessing failure mode predicted in §4.3. The evidence-grounded reward closes most of the gap, 16 to 22 points above the naive shapes across benchmarks. Layering TD-Belief GAE adds another 2 to 3 points on all 3 benchmarks. The reward shape supplies the larger share, and per-turn credit refines it.

6 Conclusion and Limitations

Long-video understanding has been framed as a model problem. We framed it as a system problem: the VLM describes what it sees, an external Bayesian controller decides where to look, and a single explicit belief between them drives routing, selection, stopping, and training. This beats prior search-based methods by up to 8.8 points using 5 to 13× fewer frames than recent RL search baselines, and the same belief is what makes the system auditable, modular, and trainable without a separate

critic. This decoupling should transfer to other tasks where perception and navigation are coupled in a single model: long-document QA, embodied search, and multi-source reasoning.

Limitations. Chunking is fixed at scene boundaries and each chunk is described in isolation, so cross-chunk evidence and ordering suffer. The SigLIP prior and the NLI scorer are also frozen, which limits how well the controller reads the question and the observations. The hypothesis h is fixed at the preview turn and committed to a single declarative claim. When that claim encodes the wrong interpretation of the question, the search runs its full budget on the wrong evidence pattern (Appendix F). Adaptive chunking, a learned controller trained against the same belief, and a multi-hypothesis variant of the scorer would address these in turn.

Acknowledgments and Disclosure of Funding

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), under the project *Perfect Recollection for Clearer Insight* (grant number EP/Y036077/1) and the computing resources and software from the AIRR Isambard-AI program.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642, 2015.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pages 273–304, 1995.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of MLLM agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246, 2025.

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.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. BED-LLM: Intelligent information gathering with LLMs and bayesian experimental design. *arXiv preprint arXiv:2508.21184*, 2025.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv preprint arXiv:2503.21776*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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*, pages 24108–24118, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13504–13514, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2024.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise RingAttention. In *The Thirteenth International Conference on Learning Representations*, 2024.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. TimeSearch-R: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. *arXiv preprint arXiv:2511.05489*, 2025.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *Synthetic Data for Computer Vision Workshop CVPR*, 2024.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. volume 37, pages 28828–28857, 2024.

Rujie Wu, Xiaojian Ma, Hai Ci, Yue Fan, Yuxuan Wang, Haozhe Zhao, Qing Li, and Yizhou Wang. LongViTU: Instruction tuning for long-form video understanding. *arXiv preprint arXiv:2501.05037*, 2025.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. volume 36, pages 76749–76771, 2023.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9127–9134, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, 2024a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2024b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2024c.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13701, 2025.

A Implementation Details

The method section specifies the system; this appendix lists the constants and code-level choices that the prose left abstract. We group them by where they appear in the pipeline and call out which values were tuned on a held-out subset versus left at sensible defaults.

Video and chunking. Videos are decoded at 2 frames per second. SigLIP-SO400M-patch14-384 encodes each frame at its native resolution, with embedding dimension $d = 1152$. A scene boundary is placed between consecutive frames when their cosine similarity drops below 0.5, and each chunk's embedding is the L2-normalized mean of its frame embeddings.

Belief constants. The prior softmax in Eq. (1) uses temperature $\tau = 1.0$. The likelihood in Eq. (3) uses evidence-strength $\beta = 1.0$. The kernel in Eq. (5) mixes temporal and semantic similarity at $\alpha = 0.5$, and its temporal bandwidth is set automatically to $\sigma = T_{\text{video}}/N$, where T_{video} is the video duration and N the chunk count. NLI logits from DeBERTa-v3-large are divided by a temperature of 3 before softmax. The temperature flattens the cross-encoder's overconfidence and keeps the log-odds scores in a range the multiplicative update can absorb. After the log-odds reduction, scores are clamped to $[-5, +5]$ to bound any single observation's influence on the belief. Figure 5 illustrates the asymmetric update from §4.1.

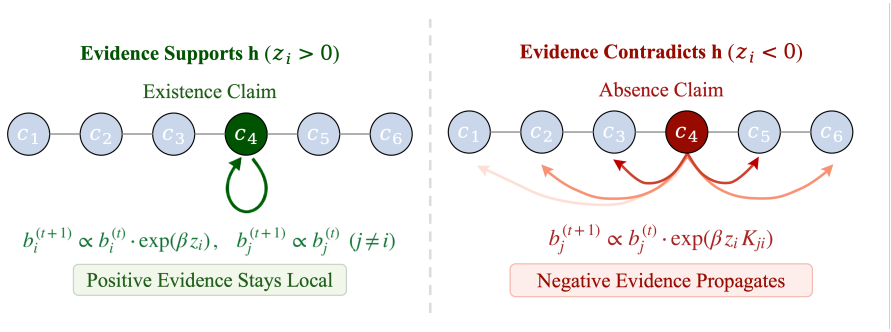


Figure 5: Asymmetric belief update. A positive observation at c_i updates only c_i (existence claim, local). A negative observation at c_i redistributes the shed mass through the kernel K_{ji} (absence claim, global).

Search budget. Each question runs for at most 8 search turns. The preview turn shows 2 low-resolution frames per chunk. Each subsequent turn shows 8 high-resolution frames sampled from the controller-chosen chunk by a DPP [Kulesza and Taskar, 2012]. To keep the prompt context bounded across turns, the actual frames from earlier search turns are stripped from the next prompt and replaced by a short text summary of the prior observations and the current belief.

Output format. Every turn produces a structured response with named tags. The controller parses these tags and ignores any free-form text outside them.

***Preview turn* (turn 1):**
- `<think>...</think>`: chain-of-thought reasoning.
- `<hypothesis>...</hypothesis>`: the declarative claim h that the NLI scorer uses for the rest of the search.
- `<summaries>{"1": "label", ...}</summaries>`: a 2-to-5-word label per chunk that the VLM uses for navigation.

***Search turns* (turns 2 through 8):**
- `<think>...</think>`: reasoning about the current chunk.
- `<observation>{"description": "...", "answer": "X"}</observation>`: a 1-to-2-sentence description of the chunk's contents. For multiple-choice questions the `"answer"` field names the option this chunk best supports, and is omitted when the chunk shows no clear evidence for any option.

***Answer turn* (any turn after the preview):**
- `<think>...</think>`: reasoning over the search trajectory.
- `<answer>...</answer>`: the final answer, which ends the search.

Routing and stopping. The routing reference count in Eq. (6) is $N_0 = 32$, the chunk count at which a single 128-frame DPP pass densely covers the video. The threshold is $\theta = 0.63$, calibrated on a

held-out subset combined across the three benchmarks. The global path uses up to 128 DPP-selected frames and a simplified prompt without belief tags. For stopping (§4.2), the noise-patience condition ends search after 5 consecutive turns with non-positive NLI scores, and the entropy-convergence condition ends search when the normalized chunk entropy $H(\mathbf{b}^{(t)})/\log_2 N$ falls below 0.7. In practice, noise patience is the more common exit. The asymmetric kernel diffuses negative evidence across chunks and keeps entropy from dropping quickly, so the entropy condition acts as a safety-net early exit when one chunk gathers very strong positive evidence rather than as the typical termination signal.

Distillation pipeline. The 26,322 SFT traces are produced by a three-stage pipeline.

(i) *Pre-filter (visual dependency).* For each multiple-choice source example, the teacher (Qwen3-VL-235B-A22B-Instruct) is shown a uniform 4-frame sample of the video and asked to answer from those frames alone. If it answers correctly from that sample, the example contains no real long-video signal and is discarded. Open-ended questions skip this check, since their free-form answers cannot be auto-verified against a single ground-truth letter.

(ii) *Trace generation.* The teacher then runs BeliefSearch on the full video in the same inference-time configuration described above (preview turn, up to 8 search turns, NLI scorer, asymmetric kernel, entropy and noise-patience stopping). The only distillation-specific addition is an *oracle hint*: after the preview, the teacher's system prompt is appended with a short instruction that names the ground-truth answer and asks the teacher to write accurate visual descriptions without revealing or hinting at it. The oracle is absent during the preview, so the declarative hypothesis h remains question-driven rather than answer-driven. The oracle text is also never written into the conversation that becomes the SFT trace, and the student model is trained on the trace alone, never on the oracle. Its only role is to anchor the teacher's per-chunk descriptions to what the frames actually show, rather than letting the teacher invent plausible-sounding but unfaithful content. Chunk selection, scoring, and stopping all run on the same controller as inference, so the trajectories the student learns from are structurally identical to the trajectories it will produce at test time.

(iii) *Post-filter.* Two checks reject low-quality traces. *Search usefulness*: the teacher's final answer must match the ground truth, the search must have run at least one observation turn, and the trace must contain a valid `<observation>` block. *Trace quality*: a trace passes if at least one observation produces a positive NLI score; if all scores are non-positive, the trace is kept only when the score sequence shows enough variation to indicate the search engaged with different chunks rather than collapsing to one repeated response. Every observation turn must also contain a `<think>` block, ensuring format compliance with the student-facing prompt.

Training. SFT runs for 3 epochs with peak learning rate 2×10^{-5}, a cosine schedule, and a short warmup. GRPO runs for 2 epochs with learning rate 5×10^{-7}, group size 4, and batch size 16. Both stages use AdamW with default $\beta_1 = 0.9$, $\beta_2 = 0.999$. The reward weights in Eq. (9) are $w_f = 0.2$ and $w_a = 2.0$. The γ and λ used by the TD-Belief GAE update are stated in §4.3. All experiments report single-seed results.

Code and dataset release. We will release the BeliefSearch codebase, the evaluation scripts, the trained checkpoint, and the full 26,322-trace SFT dataset under a permissive license upon acceptance.

B Hypothesis Generation

The preview turn produces a single declarative hypothesis h that the NLI scorer uses for the remainder of the search (§3). The hypothesis is the system's translation of the question into something the scorer can reason about: it names the visible evidence pattern the controller should look for in each chunk, without committing to a specific answer. Getting this translation right matters, because every observation is scored against the same fixed reference. This appendix walks through the choices behind it.

Why declarative. NLI cross-encoders are trained on pairs of declarative statements drawn from SNLI and MultiNLI [Bowman et al., 2015, Williams et al., 2018]. Feeding an interrogative question as the hypothesis pushes the scorer outside its training distribution: "does this observation entail the question *'How many valid goals does Jesse score?'*?" is not a meaningful entailment query, and the resulting probabilities become noisy. Rephrasing the question as a declarative claim restores the input

format NLI was trained on, and the cross-encoder produces entailment probabilities that track the semantic relation between an observation and the claim rather than the surface form of the question.

Why the VLM at the preview turn. A rule-based question-to-statement transformation handles simple cases but breaks on negation, multi-clause phrasing, and domain-specific vocabulary. Robust handling of all three already lives in the VLM's language head, so format conversion is a capability the system uses without adding a separate component. The preview turn is the right place to do this for a second reason as well: by the time the VLM is asked to write h, it has already seen low-resolution frames from every chunk and has a rough sense of what the video is about. The hypothesis can therefore name objects, settings, or actions that actually appear in the video rather than abstract entities the question alone cannot anchor. Running the rephrasing once at the preview turn also keeps h stable across the search and avoids an extra VLM call on every observation.

What a good hypothesis looks like. A useful hypothesis describes the kind of visible evidence that would resolve the question. The exact form depends on the question.

- For *counting* or *summative* questions, the hypothesis names the evidence pattern rather than the count itself: "How many valid goals does Jesse score?" becomes "a basketball goes into the hoop and a score counter shows the result." The controller collects chunks that entail this pattern, and the VLM tallies the goals at answer time.
- For *cause* or *inference* questions, the hypothesis fixes a specific candidate explanation: "What can be inferred as the reason for the surge in popularity?" becomes "social media promotion is visible." Positive evidence confirms this candidate, and if every chunk returns a negative score the kernel diffuses mass elsewhere so the VLM can fall back on an alternative inference at answer time.
- For *identity* or *factual* questions, the hypothesis is essentially a candidate answer paired with what it would look like on screen: "Which media outlet reported on this video?" becomes "NBC NEWS logo is visible." This works when SigLIP can localize the visual evidence and breaks when it cannot: small in-frame text such as logos, watermarks, and captions falls outside what scene-level SigLIP encodes well.

Robustness. A hypothesis that no chunk entails fails gracefully: every observation returns a non-positive NLI score, the noise-patience condition ends the search early, and the VLM answers from whatever evidence the preview turn provided. The harder case is a hypothesis that does match a chunk but encodes the wrong interpretation of the question. Examples 8 and 9 in Appendix F show this failure mode in detail, and the closing analysis there discusses architectural remedies (a learned scorer or a multi-hypothesis variant) within the same Bayesian framework.

C Training Stages

This appendix supplies the numbers behind the training-dynamics paragraph in §5.2. Table 6 traces the model through four points in the training pipeline. The controller, the prompts, the search loop, and the routing rule are held fixed across rows, so the differences isolate what each training stage contributes to the VLM's descriptions, separately from any change in how the controller navigates.

Table 6: Training pipeline progression on Qwen3-VL-8B.

Stage	VME-S	VME-M	VME-L	**VME**	**LVB**	**MLVU**
Base model (128f, no search)	80.9	64.9	62.7	69.5	66.0	67.9
+ Zero-shot BeliefSearch	75.9	24.2	9.0	36.4	33.2	58.9
+ SFT	73.0	62.7	54.7	63.5	60.7	64.5
+ GRPO (TD-Belief GAE)	79.8	69.6	67.4	**72.3**	**66.9**	**72.0**

Base model. A single VLM call on 128 DPP-selected frames, no search loop, no fine-tuning. This is the strongest single-pass reference available to the same VLM and the bar the multi-turn system must clear to justify its complexity.

Zero-shot BeliefSearch. The same base VLM is dropped into our multi-turn protocol with no fine-tuning. Accuracy collapses unevenly: a 5-point drop on short videos but more than 30 points on the medium and long splits, and around 33 points overall on VME and LVB. The pattern reflects three

things the base model has never been asked to do: rephrase a question into a declarative hypothesis, emit observations in the structured <think>/<observation>/<answer> format, and describe each chunk in a way the scorer can act on. Short videos route to the global path more often (Table 3) and bypass most of the protocol, so they collapse less. Long videos always go through the search path, hit all three failure modes at once, and accuracy bottoms out near zero on the longest split.

SFT. Training on the 26,322 distilled traces (Appendix A) recovers most of the gap to the base model. The model regains 5 to 27 points across the three benchmarks relative to zero-shot, and the long-video pattern reverses: VME-L returns to 54.7%, well above the zero-shot 9.0%. What SFT teaches at this stage is the *format and protocol*: the model learns when to write a hypothesis, what an observation looks like, and roughly how to describe the contents of a chunk. What it does not yet teach is *description quality* measured against the scorer. The descriptions imitate the teacher's surface form but remain generic enough that the scorer often cannot extract strong entailment or contradiction from them. The result is a working multi-turn run that still sits a few points below the single-pass baseline.

GRPO with TD-Belief GAE. The RL stage clears the single-pass baseline on every benchmark, by 2.8, 0.9, and 4.1 points on VME, LVB, and MLVU respectively. The improvement is uneven by length: short videos are roughly tied with the base model (VME-S 79.8 vs 80.9) because most short questions route to the global path and bypass the search loop, while the medium and long splits each gain 4.7 points, where the search runs the most turns and per-turn description quality matters most. The controller, the prompts, and the search hyperparameters are unchanged from the SFT stage, so the gain reflects the change in what the VLM writes at each turn rather than any change in how the controller navigates. The reward credits the model only when its trajectory's belief entropy actually drops (Eq. 9), and the per-turn GAE advantage attributes that drop to specific turns (Eq. 8). Both signals push the model toward the same behavior: writing descriptions the scorer can convert into a belief update.

D Training Approaches Compared

TD-Belief Learning sits in a small space of options for credit assignment in multi-turn RL. Table 7 compares it to two standard reinforcement-learning baselines (GRPO, PPO) and to two video-search systems that train the navigation policy through reinforcement learning (TimeSearch-R, Video-R1).

Table 7: Training approaches compared.

Approach	Per-turn credit	Critic	Reward (main)
TimeSearch-R	No	None	Completeness self-check (extra LLM call)
Video-R1	No	None	Temporal contrast
GRPO	No	None	Task-specific
PPO	Yes (GAE)	Learned $V(s)$	Task-specific
TD-Belief (ours)	**Yes (GAE)**	**Belief entropy**	**Acc** $\cdot(1 - H_T/H_0)$

The axis that matters most for our setting is per-turn credit. GRPO is a bandit method: each rollout earns a single reward, and the group-relative baseline assigns the same advantage to every turn in the trajectory, so the model cannot tell which turn carried the gain. PPO recovers per-turn credit through a learned value function, which restores temporal resolution but adds a separate network whose target is sparse episode-level rewards. That target is hard to fit cleanly, especially for trajectories of variable length. TimeSearch-R and Video-R1 both stay closer to GRPO. Their reward shapes are task-specific (a completeness self-check that costs an extra LLM call, and a temporal contrast term respectively), but the rewards still apply to the trajectory as a whole and there is no critic. TD-Belief Learning takes neither route. The belief is a state the controller already maintains for inference, and its entropy is an analytical proxy for value (§4.3), so the per-turn TD signal comes for free without training a new network or designing a new reward shape.

E MLVU by Task Type

MLVU contains seven task categories, and they distribute evidence across the video in different ways. The ablation in §5.2 reports a 7.7-point gain of BeliefSearch over random chunk selection on the

benchmark as a whole. This appendix decomposes that gain by category, so we can check that the improvement is not concentrated on a few task types.

Table 8: MLVU per-task accuracy: BeliefSearch vs. random chunk selection. The seven task categories cover 2,170 of MLVU dev's 2,174 examples; the remaining 4 lack a clean task label and are excluded from this breakdown.

Task type	Count	BeliefSearch	Random	Δ
Order reasoning	259	48.7	41.3	+7.4
Anomaly recognition	200	58.6	51.0	+7.6
Counting	206	48.3	41.7	+6.6
Topic reasoning	262	92.9	82.4	+10.5
Needle QA	355	85.4	76.1	+9.3
Plot QA	536	80.5	71.6	+8.9
Ego reasoning	352	68.8	65.3	+3.5
Overall	2170	72.0	64.3	+7.7

The largest gains are on topic reasoning, needle QA, and plot QA, between 8.9 and 10.5 points each. In all three, the answer evidence concentrates in a small handful of moments, which is exactly where a belief that can locate those moments has the most to add. Counting, anomaly recognition, and order reasoning sit in the middle of the table at 6.6 to 7.6 points, where the evidence is moderately localized but still benefits from focused observation. Ego reasoning is the opposite extreme. A long stretch of first-person footage spreads the answer evidence across many moments rather than placing it in a few, so focused selection has less to recover, and the gain shrinks to 3.5 points. The improvement is positive on every category nonetheless, and no task type favors random or sequential selection over the belief-guided rule.

F Example Runs

Section 5.2 sketches three trajectories from the qualitative figure. This appendix walks through nine more examples across all three benchmarks, organized into four themes: *answer flipping under conflicting evidence*, *both baselines fail*, *methodical elimination*, and *failure cases*. Each example contrasts BeliefSearch (belief-guided multi-turn search), TimeSearch-R (model-driven tool calls), and Video-R1 (a single 32-frame uniform pass). All scores reported below are NLI log-odds (Eq. (2)), and all entropies are in bits. Both baselines occasionally state one letter inside their reasoning and then emit a different letter in the `<answer>` tag. Following standard evaluation, we always report the letter from the `<answer>` tag, which can therefore differ from what the verbatim reasoning above it concludes.

F.1 Answer flipping under conflicting evidence

Two examples where the search cycles through several candidate answers before a late positive observation locks in the correct one. The first example commits to a single wrong answer and recovers on the final positive observation; the second example tries three different answers in succession before settling. Both baselines commit early to a locally plausible answer and never recover.

Example 1 (`videomme_565-3`). VME, 8m33s, 6 turns, $\Delta H = +1.25$ bits.
Q: What mode of transport do the people in the video take to get to Horseshoe Bay?
Options: (A) Helicopter (B) Car (C) Bus (D) Boat **GT:** A
Outcomes: BeliefSearch **A** ✓ · TimeSearch-R **B** × · Video-R1 **B** ×

BeliefSearch

Hypothesis h: A helicopter is visible.

T1, c5. Chunk 5 shows a car interior with a person driving and a view of a desert road with red rock formations. The frames include a hand on the steering wheel and a side mirror view of the road, indicating the vehicle is in motion. This supports the hypothesis that a car is being used for transport. The presence of the car and the road suggests driving is the mode of transport, which aligns with option B. I will describe the visible car and road without mentioning the location or other details.

Observation: "A person is driving a car on a desert road with red rock formations visible outside." -2.15, ans $\rightarrow$ B

T2, c7. The frames show a group of people at Horseshoe Bend, with one person in a red cap and backpack, and others nearby, including a man in a hat and sunglasses. A small dog is visible on a leash, and the background shows a desert landscape with red rock formations. There is no visible helicopter, car, bus, or boat in these frames — only people and a dog. Since the question is about the mode of transport to Horseshoe Bay, and no transport is visible here, I cannot assign an answer yet. I will continue searching.

Observation: "A group of people and a small dog are standing at Horseshoe Bend with red rock formations in the background." -2.62

T3, c2. The frames from Chunk 2 show a person driving a car on a desert road with red rock formations visible outside, and another person in the passenger seat gesturing while wearing sunglasses. The car is clearly in motion, and the setting matches the route to Horseshoe Bay. This directly supports option B (Car) as the mode of transport. The belief score for Chunk 2 was 0.11, and after searching Chunk 5 and 7, the belief dropped, but Chunk 2's visual evidence is strong. I will describe the car and people inside, confirming the transport mode.

Observation: "A person is driving a car on a desert road with red rock formations visible outside, and another person in the passenger seat is gesturing while wearing sunglasses." -2.82, ans $\rightarrow$ B

T4, c4. The current chunk (4) shows a person standing near a white car with a rooftop tent, in a forested area, suggesting they are setting up camp. The car is parked, and the person is interacting with it, but there is no visible movement toward Horseshoe Bay or any indication of transport mode. The frames do not show a helicopter, boat, or bus. Since the question is about the mode of transport to Horseshoe Bay, and this chunk only shows a car parked at a campsite, it does not provide evidence for any option. I will describe what is present: a person near a car with a rooftop tent in a forest. No answer letter is included because nothing relevant to the transport mode is visible.

Observation: "A person in an orange blanket stands next to a white car with a rooftop tent in a forested area, with a ladder leaning against the car." -2.61

T5, c9. The frames from Chunk 9 show a yellow helicopter on a landing pad with people nearby, including a man in a black shirt and a woman in a white top and hat. The helicopter is stationary, and people are interacting with it, suggesting it is the mode of transport used to reach the location. The presence of the helicopter and the people's proximity to it strongly indicate that the transport to Horseshoe Bay is by helicopter. This matches option A.

Observation: "A yellow helicopter is on a landing pad with people standing nearby, including a man in a black shirt and a woman in a white top and hat, suggesting it is the mode of transport." $+1.93$, ans $\rightarrow$ A

Final answer: A ✓

TimeSearch-R

Call 1. To answer the question, I need to identify the mode of transport used by the people in the video to reach Horseshoe Bay. I will search for scenes related to Horseshoe Bay and the mode of transport used to get there.

Query: "Horseshoe Bay" *Range:* 0.0–513.2 s, 8 frames.

Call 2. The frames provided show various locations and activities, but the specific frame at 403.7s clearly indicates "Horseshoe Bend" as the location. The question asks about the mode of transport to Horseshoe Bend. I need to focus on the frames related to Horseshoe Bend to determine the mode of transport.

Query: "Horseshoe Bend" *Range:* 300.0–513.2 s, 8 frames.

Closing reasoning. The frames clearly indicate that the people in the video are at Horseshoe Bend and are traveling by car. The frames show them in a car, suggesting that the mode of transport to Horseshoe Bend is by car. Therefore, I can conclude that the mode of transport used to get to Horseshoe Bend is a car.

Final answer: B ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The question is asking about the mode of transport used by the people in the video to get to Horseshoe Bay. Now, let's break it down.

First, I need to consider the options provided: A. Helicopter - This is a possibility, but there's no clear indication in the video that they used a helicopter. B. Car - This is a common mode of transport for short distances, and it's often used for sightseeing trips. C. Bus - While buses can be used for long-distance travel, there's no evidence in the video that they used a bus. D. Boat - This is unlikely, as there's no water body visible in the video, and the landscape is desert-like.

Now, let's look at the video. The people are seen standing in front of a car, and they appear to be preparing for a trip. They are also seen driving a car, which suggests that they used a car to get to their destination.

Hmm, it seems that the most logical and common mode of transport for the people in the video to get to Horseshoe Bay is a car. The video does not show any evidence of them using a helicopter, bus, or boat.

Oh, I see. The most reasonable answer based on the video and the options provided is that they used a car.

Let's verify this one last time. The video shows the people driving a car, and there's no indication of them using any other mode of transport. Therefore, the answer is clear.

Final answer: B

Let's break it down one last time to ensure. The video shows the people driving a car, and there's no evidence of them using a helicopter, bus, or boat. Therefore, the answer is B.

Final answer: **B** ×

Example 2 (`videomme_529-2`). VME, 14m16s, 6 turns, $\Delta H = +0.85$ bits.
Q: How many kinds of earrings are made in this video?
Options: (A) 2 (B) 3 (C) 4 (D) 5 **GT:** C
Outcomes: BeliefSearch **C** ✓ · TimeSearch-R **B** × · Video-R1 **A** ×

BeliefSearch

Hypothesis h: Multiple distinct earring designs are visible.

T1, c12. In Chunk 12, the person is shown wearing and adjusting two distinct earring designs: one is a pink heart-shaped pendant on a gold chain, and the other is a small silver charm with a blue planet-like bead. These are clearly different in shape and color, confirming at least two distinct earring types. I need to check if more are visible in other chunks to determine the total count. I will now examine Chunk 11, which has the next highest belief and shows the person wearing the pink heart earring, to see if it's the same or if another type is present.

Observation: "A person is wearing a pink heart-shaped earring on a gold chain and adjusting it with their hand." -1.46, ans $\rightarrow$ A

T2, c2. The frames show hands handling different types of earring components: a package of fishhook earrings, a package of eye pins, and a package of small silver beads. The person is sorting and placing these into a clear organizer, indicating they are preparing to make multiple earring types. The pink heart earring seen in Chunk 12 is one type, and the leaf charm in Chunk 8 is another. The small silver beads and eye pins suggest a third type. I have not yet seen a fourth distinct earring design. I will search Chunk 8 next to confirm the leaf charm is being made into an earring.

Observation: "Hands sort and organize fishhook earrings, eye pins, and small silver beads into a clear container, preparing to assemble multiple earring types." $+1.49$, ans $\rightarrow$ B

T3, c3. The current chunk shows hands selecting and handling small silver leaf-shaped charms from a compartmentalized organizer, alongside other earring components like fishhooks and eye pins. The person is actively assembling or preparing these charms, which appear to be intended as earring backs or pendants. This is a distinct earring type from the pink heart earring seen earlier, supporting the count of at least two different earring designs. I will now search Chunk 8, which is labeled "Leaf charm," to confirm if it's being made into an earring and whether it's a third distinct type.

Observation: "Hands select and handle small silver leaf-shaped charms from a compartmentalized organizer, preparing them for assembly as earring components." -2.74, ans $\rightarrow$ B

T4, c8. The frames show hands assembling two distinct earring types: one with pink heart-shaped pendants on gold chains and another with silver leaf-shaped charms on silver chains. The pink heart earring is shown being worn and adjusted, confirming it as a finished item. The leaf charm is also shown being attached to a chain, indicating it's being made into an earring. These are two distinct designs. I need to check if a third type is visible — perhaps the small silver bead or another component. I will now examine Chunk 7, which has the highest belief and is labeled "Pink heart earring," to confirm if it shows the same or a different earring type.

Observation: "A hand holds a pink heart-shaped earring on a gold chain, then later holds a silver leaf-shaped earring on a silver chain, showing two distinct earring designs being assembled." $+1.92$, ans $\rightarrow$ A

T5, c7. The current belief is highest for Chunk 8 (0.44), which already showed two distinct earring designs: pink heart and silver leaf. I need to check if a third type is visible. Chunk 7 shows hands assembling a pink heart earring on a gold chain, then later holding a silver leaf earring on a silver chain — confirming two types. No other earring types are visible in this chunk. Since the question asks for "kinds of earrings made," and I've seen two distinct designs being assembled, I need to check if a third is present elsewhere. Chunk 10 shows a small silver bead being handled — possibly a third earring type. I will search Chunk 10 next to confirm if it's a distinct earring design.

Observation: "A hand holds a small silver bead and uses pliers to attach it to a gold chain, forming a third distinct earring design." $+1.91$, ans $\rightarrow$ C

Final answer: **C** ✓

TimeSearch-R

Call 1. To determine how many kinds of earrings are made in the video, I need to identify all the earrings that are created. I will review the video frames to identify distinct types of earrings being crafted.

Query: "earrings being made" *Range:* 0.0–856.3 s, 8 frames.

Closing reasoning. From the provided frames, I can see that there are different types of earrings being made. Based on the frames, I will count the distinct types of earrings.

Final answer: **B** ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The video shows a person making earrings, and I need to determine how many kinds of earrings are made.

First, let's break it down. The person is seen working with beads and wires, and there are different types of beads and wires used. The earrings appear to have different designs and components, which suggests multiple kinds.

Now, let's consider the options: A. 2 - This would mean only two kinds of earrings are made. B. 3 - This would mean three kinds of earrings are made. C. 4 - This would mean four kinds of earrings are made. D. 5 - This would mean five kinds of earrings are made.

Hmm, it seems like the person is making a variety of earrings, each with its own unique design. The presence of different beads and wires, along with the different components used, suggests that there are indeed multiple kinds of earrings.

Oh, I see. The person is likely making a few different kinds of earrings, each with its own unique design. This aligns with the option that suggests more than two kinds of earrings.

Let's verify this. The person is seen working with different beads and wires, and the earrings have different designs. This confirms that there are indeed multiple kinds of earrings made.

After all this reflection, I am confident that the correct answer is B, as it aligns with the variety of earrings made in the video.

Let's break it down one last time to ensure. The person is making a few different kinds of earrings, each with its own unique design, which confirms the answer.

Final answer: B

Let's double-check the options one last time to ensure there are no mistakes. The answer B is the most logical and aligns with the visual evidence in the video.

Final answer: A ×

F.2 Both baselines fail

Three examples where neither single-pass nor model-driven search reaches the right answer, but belief-guided search converges. Together they cover a short MLVU clip, a short LVB clip, and a 42-minute VME clip, so the pattern is not specific to one benchmark or duration.

Example 3 (`mlvu_needle_Q9_needle_36.mp4`). MLVU, 7m58s, 3 turns, $\Delta H = +1.35$ bits.
Q: What type of view is provided of the tropical beach in the video?
Options: (A) Close-up view (B) Bottom-up view (C) Top-down aerial view (D) Side view
GT: C
Outcomes: BeliefSearch C ✓ · TimeSearch-R A × · Video-R1 A ×

BeliefSearch

Hypothesis h: A top down aerial view of the beach is visible.

T1, c1. The current chunk (Chunk 1) shows close-ups of a cat's face and a person's legs in a room, with no beach or aerial perspective visible. Since the question is about the beach's view, and this chunk contains no relevant visual evidence, I will not assign an answer letter. I will continue searching higher-belief chunks that may contain the beach view.

Observation: "Close-up of a cat's face and a person's legs in a room with a radiator and doorway visible." −4.34

T2, c6. The frames from Chunk 6 show a clear overhead, wide-angle perspective of a tropical beach with turquoise water, white waves, rocks, and green vegetation along the shore. This is consistent with a top-down aerial view, which matches option (C). The camera angle is high and directly above the scene, capturing the entire beach area without any side or bottom-up perspective. Since this chunk provides direct visual evidence of the beach from an aerial viewpoint, I can confidently answer the question.

Observation: "A top-down aerial view of a tropical beach with turquoise water, white waves, rocks, and green vegetation along the shore." +1.80, ans → C

Final answer: C ✓

TimeSearch-R

Call 1. To answer the question about the type of view provided of the tropical beach in the video, I need to identify the relevant frames that show the beach. I will search for frames that depict a tropical beach using the seek_video_frames tool. The query will be "tropical beach" to find the frames that showcase this environment.

Query: "tropical beach" *Range:* 0.0–478.7 s, 8 frames.

Closing reasoning. The frames provided clearly depict a tropical beach scene with aerial views of the coastline, turquoise waters, and rocky formations. These frames align with the query "tropical beach," confirming that the video provides an aerial view of a tropical beach.

Final answer: A ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The question is asking about the type of view provided of the tropical beach in the video. Now, let's break it down.

First, I need to recall the details of the video. The video shows a tropical beach with clear blue water, white sand, and lush greenery. The beach is surrounded by rocks and trees, and there are people walking along the shore. The view is from above, providing a bird's-eye perspective of the beach.

Now, let's consider the options: A. Aerial view - This would be a view from above, looking down at the beach. B. Ground-level view - This would be a view from the ground, looking up at the beach. C. Close-up view - This would be a view focusing on a specific part of the beach, such as a person or a rock. D. Side view - This would be a view from the side, showing the beach from a horizontal angle.

Hmm, the video clearly shows a bird's-eye perspective of the beach, which is characteristic of an aerial view. The camera is not at ground level, not focusing on a specific part of the beach, and not showing a side view. It is definitely from above, providing a comprehensive view of the beach and its surroundings.

Oh, I see. The correct answer is the aerial view. The video provides a bird's-eye perspective of the tropical beach, which is the most comprehensive and detailed view possible from the given options.

Let's verify this one last time. Yes, the video is indeed an aerial view, providing a clear and detailed look at the beach and its features.

So, the correct answer is A.

Let's break it down one more time to ensure. The video is not a ground-level view, not a close-up view, and not a side view. It is an aerial view, which is the only option that fits the description.

Final answer: A ×

Example 4 (`X0U3fP0tZyY_0`). LVB, 7m38s, 3 turns, $\Delta H = +1.12$ bits.
Q: A man wearing a red lettered t-shirt is explaining in front of a wall of string instruments. What is the first object that appears after mentioning "performed by a wonderful cuatrista, Fabiola Mendez"?
Options: (A) woman in a black dress playing the piano (B) woman in black clothing playing a small string instrument (C) woman in purple clothing playing a small string instrument (D) woman in a white dress playing the piano **GT:** B
Outcomes: BeliefSearch **B** ✓ · TimeSearch-R **none** × · Video-R1 **C** ×

BeliefSearch

Hypothesis h: A woman playing a small string instrument is visible.

T1, c5. The frames in Chunk 5 show a woman with glasses and dark curly hair playing a small string instrument (cuatro) while seated. She is wearing a black top and is actively playing the instrument, with a man in a red-lettered shirt nearby. The text overlay identifies her as "Fabiola Mendez, Cuatrista," confirming she is the "wonderful cuatrista" mentioned in the question. The first object that appears after her mention is her playing the instrument, which matches option B. I can now answer confidently.

Observation: "A woman in black clothing is sitting and playing a small string instrument while a man in a red-lettered shirt stands nearby." $+1.84$, ans $\rightarrow$ B

T2, c3. The search history indicates that Chunk 5 already showed a woman in black clothing playing a small string instrument, which aligns with option B. The current chunk (Chunk 3) shows the man in the red-lettered shirt gesturing in front of the wall with string instruments, but no woman is visible here. Since the question asks for the first object after the mention of Fabiola Mendez, and Chunk 5 already provided evidence for option B, I should confirm if this is the first appearance. However, the system's belief update suggests Chunk 5 is the key. I will now search Chunk 4, which has the next highest belief, to see if it shows the woman or any transition.

Observation: "A man in a black t-shirt with red lettering gestures in front of a wall with wooden string instruments hanging on it." -2.56, ans $\rightarrow$ B

Final answer: **B** ✓

TimeSearch-R

Call 1. To answer this question, I need to identify the first object that appears after the text "performed by a wonderful cuatrista, Fabiola Mendez." in the video. I will search for the video frames around the mentioned text to pinpoint the first object that appears visually.

Query: "first object after 'performed by a wonderful cuatrista, Fabiola Mendez."' *Range:* 60.0–65.0 s, 8 frames.

Closing reasoning. The first object that appears after the text "performed by a wonderful cuatrista, Fabiola Mendez." is a guitar. Based on the frames provided, the first object that appears visually is a guitar.

Final answer: **none** ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this carefully. The question is asking about the first object that appears after mentioning 'performed by a wonderful cuatrista, Fabiola Mendez.' The context provided is a man explaining in front of a wall with musical instruments, and the question is about the sequence of objects after this mention.

First, let's break it down. The man is explaining, and there's a wall with musical instruments. The mention of Fabiola Mendez is likely a reference to a specific instrument or performance. The question is asking about the first object that appears after this mention.

Now, let's consider the options. The options are A, B, C, and D. The question doesn't specify which object appears first, so we need to infer based on the context.

Hmm, the context suggests that the first object after the mention of Fabiola Mendez is likely the instrument she performed on. The question doesn't specify the exact object, but it's reasonable to assume it's the first object in the sequence after the mention.

Oh, I see. The first object that appears after mentioning Fabiola Mendez is likely the instrument she performed on. This is the most logical sequence based on the context provided.

Let's verify this. The man is explaining, and the wall has instruments. The mention of Fabiola Mendez is a reference to a specific instrument. The first object after this mention is likely the instrument she performed on.

So, the first object that appears after mentioning 'performed by a wonderful cuatrista, Fabiola Mendez.' is the instrument she performed on.

Final answer: C

Let's break it down one more time to ensure. The first object after the mention of Fabiola Mendez is the instrument she performed on. This is the most logical sequence based on the context.

Final answer: **C** ×

Example 5 (`videomme_718-1`). VME, 42m25s, 6 turns, $\Delta H = +0.94$ bits.
Q: Based on the early part of the video, what can we infer as the reason for the sudden surge in popularity of crystals and gemstones?
Options: (A) New variety of crystals has emerged (B) Enhanced accessibility and affordability (C) Social media promotion (D) Growing interest in holistic wellness **GT:** C
Outcomes: BeliefSearch **C** ✓ · TimeSearch-R **D** × · Video-R1 **D** ×

BeliefSearch

Hypothesis h: Social media promotion is visible.

T1, c1. The frames from Chunk 1 show a person holding a purple crystal near their face, smiling, and interacting with it in a way that suggests personal use or demonstration. The presence of a tattooed arm and a ring indicates a casual, personal context. The background includes a lamp and a wall, suggesting an indoor setting, possibly for a video or photo. The person's expression and the way they hold the crystal imply a personal or wellness-related use, which aligns with holistic wellness. The phone in Chunk 2 shows a photo of the crystal, indicating social media sharing. The article in Chunk 6 mentions "healing crystals" and "wellness craze," reinforcing the wellness angle. The hypothesis is that social media promotion is visible, but the frames show personal interaction and display, not direct promotion. The best evidence is the person's personal use and the phone photo, which supports social media promotion. I will describe the visible actions: person holding crystal, smiling, and phone photo.

Observation: "A person holds a purple crystal near their face and smiles, with a phone displaying a photo of the crystal, suggesting personal use and social media sharing." $+1.85$, ans $\rightarrow$ C

T2, c7. The current chunk (7) shows a person crouching outdoors near large white crystals, handling them with bare hands, and a close-up of a foot near the crystals. This suggests direct interaction with raw or semi-processed crystals in a natural or outdoor setting, which aligns with the idea of collecting or preparing them for use. While this doesn't directly show social media or wellness branding, it supports the idea of personal engagement with crystals, which could be part of a broader trend. However, the strongest evidence so far (from chunk 1) was a person smiling with a crystal and a phone displaying a photo — which strongly implies social media promotion. Since the question is about inference from the early part, and chunk 1 already provided strong evidence for option C, I should now check if any other chunk provides conflicting or stronger evidence. Chunk 6 showed an article mentioning "healing crystals" and "wellness craze," which supports D. But chunk 1's phone and smile are more direct evidence of promotion. I will now search chunk 6 to confirm if the article explicitly links to social media or wellness.

Observation: "A person crouches outdoors near large white crystals, handling them with bare hands, and a close-up shows a foot near the crystals, suggesting direct interaction in a natural setting." -2.06, ans $\rightarrow$ D

T3, c4. The current chunk shows a person with green hair holding a small clear crystal, smiling and speaking, with a lit candle and other crystals in the background. This suggests a personal, possibly promotional or educational context, consistent with social media promotion. The presence of a phone in earlier chunks and the direct engagement with the camera support this. I will describe the visible actions and objects without inferring absence. The description focuses on the person, the crystal, and the setting. Since the question is about inference from early video, and this chunk shows active handling and display, it supports option C. I will output the observation with answer C.

Observation: "A person with green hair holds a small clear crystal while smiling and speaking, with a lit candle and other crystals visible in the background." -2.03, ans $\rightarrow$ C

T4, c9. The current belief is highest for Chunk 11 (0.10), which shows a person holding a pink crystal with a smile and a phone displaying a photo — strong evidence of social media promotion. Chunk 10 (0.09) shows a man with a mic and a woman holding a phone, also suggesting recording or sharing. Chunk 8 (0.08) shows people outdoors with crystals, possibly in a group setting. The most relevant chunk for the question is Chunk 11, as it directly shows a person smiling while holding a crystal and a phone displaying a photo, which implies sharing or promotion. I will search Chunk 11 next to confirm if the phone is being used to share the crystal image, which would support option C. If not, I will check Chunk 10 for similar evidence.

Observation: "A person holds a pink crystal and smiles while a phone displays a photo of the crystal, suggesting social media sharing." $+1.82$, ans $\rightarrow$ C

T5, c11. The current belief is highest for Chunk 9 (0.39) and Chunk 1 (0.26), both showing people holding crystals with phones displaying photos — strong evidence for social media promotion. Chunk 11 shows a man at a booth with a microphone and a sign reading "MINERAL & GEM," and people walking past tents with crystals on display, suggesting a public event or market. This supports accessibility and affordability (B) as people are freely browsing and purchasing. The presence of a microphone and event signage implies organized promotion, which could tie to social media or wellness, but the most direct visual is the public display and interaction. I will describe the visible elements: the man with microphone, the sign, and people near the display. This best supports B.

Observation: "A man stands at a booth with a microphone and a sign reading 'MINERAL & GEM' while people walk past tents displaying crystals, indicating a public event with accessible displays." -2.11, ans $\rightarrow$ B

Final answer: **C** ✓

TimeSearch-R

Call 1. To answer this question, I need to analyze the early part of the video to identify the main reasons for the surge in popularity of crystals and gemstones. The video frames suggest a focus on the beauty industry and the use of crystals in wellness practices. I will search for frames that highlight social media promotion and the wellness craze.

Query: "crystals and gemstones social media promotion" *Range:* 0.0–500.0 s, 8 frames.

Closing reasoning. The frames clearly indicate that social media promotion has played a significant role in the surge of interest in crystals and gemstones. The repeated mentions of TikTok and YouTube content, along with the frames showing people taking photos of crystals and sharing them on social media, strongly suggest that social media promotion has been a key factor.

Final answer: D ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The question is asking about the reason for the sudden surge in popularity of crystals and gemstones. The options provided give us a range of possibilities, and I need to consider each one carefully.

First, let's break it down. The video shows various scenes of crystals and gemstones, and there's a mention of a woman discussing a beauty routine, which might be related to the use of crystals in holistic wellness. This suggests that there might be a growing interest in holistic wellness and alternative spirituality.

Now, let's consider the other options: A. A whole new variety of crystals and gemstones has emerged. While it's possible that new varieties have emerged, there's no direct evidence in the video to support this. B. Enhanced accessibility and affordability of gemstones. The video doesn't provide any information about the accessibility or affordability of gemstones. C. Social media promotion. The video does show a woman discussing a beauty routine, which could be related to social media promotion, but there's no direct evidence of this. D. Growing interest in holistic wellness and alternative spirituality. This seems the most plausible based on the scenes in the video and the mention of holistic wellness.

Hmm, it seems that the most logical reason for the surge in popularity of crystals and gemstones is the growing interest in holistic wellness and alternative spirituality. The video shows various scenes related to crystals and gemstones, and there's a mention of a woman discussing a beauty routine, which aligns with this interest.

Oh, I see. The video provides a strong indication that the growing interest in holistic wellness and alternative spirituality is the reason for the surge in popularity of crystals and gemstones.

Let's verify this one last time. The video does not provide evidence for the other options, and the scenes related to crystals and gemstones align with the interest in holistic wellness.

Final answer: D

Let's break it down one last time to ensure. The video does not provide evidence for A, B, or C, and the scenes related to crystals and gemstones align with D.

Final answer: D ×

F.3 Methodical elimination

Two examples where the search rules out several chunks before locating the right one. The negative observations matter: each pushes mass through the asymmetric kernel and concentrates belief on the chunks that have not yet been shown to be wrong.

Example 6 (`mlvu_needle_Q58_needle_50.mp4`). MLVU, 7m38s, 6 turns, $\Delta H = +1.35$ bits.
Q: What are the two kids doing in the paddy field?
Options: (A) Studying (B) Sleeping (C) Having fun (D) Playing football **GT:** C
Outcomes: BeliefSearch C ✓ · TimeSearch-R A × · Video-R1 A ×

BeliefSearch

Hypothesis h: Two kids are playing.

T1, c6. Chunk 6 shows underwater scenes with fish and plants, then a blurry transition to a sunset with palm trees. No kids are visible. The frames do not show any activity related to the question. I will continue searching.

Observation: "Underwater fish swimming among green plants, then a sunset with palm tree silhouettes." -3.45

T2, c1. The current chunk (Chunk 1) shows a close-up of a hippopotamus in a grassy field, then underwater views with fish and reeds, and finally an aerial shot of a wetland with birds and water. No kids are visible in any of these frames. The question is about two kids in a paddy field, but this chunk contains only animals and water — no human activity. Since the correct answer is C (having fun), and no kids are visible here, this chunk does not support any option. I will describe what is present: a hippo, fish, and birds. No answer letter.

Observation: "A hippopotamus is visible in a grassy field, followed by underwater fish and reeds, then birds flying over a wetland." -3.58

T3, c8. Chunk 8 shows a close-up of a bird with an orange beak and two small chicks on sandy ground. The bird appears to be feeding or interacting with the chicks. This is not two kids, but it is a pair of young animals being cared for — which may loosely relate to "having fun" if interpreted as playful interaction. However, since the question specifies "two kids," and no human children are visible, this does not directly support any option. I will describe what is present without inferring relevance. The frames show birds and chicks, not kids. I will not assign an answer yet.

Observation: "A bird with an orange beak is near two small chicks on sandy ground, appearing to interact with them." -3.06

T4, c5. The frames from Chunk 5 show two children in a grassy field with a cow nearby. One child is wearing a purple top and the other a blue top; they are standing and moving around, not engaged in studying or sleeping. Their posture and proximity to each other suggest playful interaction. The presence of a cow and open field supports the idea of leisure activity. This aligns with option (C) "Having fun" — not football (no ball or kicking motion), not studying (no books or focused activity). I will describe the visible behavior without referencing location or clothing details beyond what's necessary.

Observation: "Two children are standing and moving in a grassy field with a cow nearby, appearing to interact playfully." $+0.95$, ans $\rightarrow$ C

T5, c3. The current belief is highest for Chunk 5 (0.56), which showed two children interacting playfully with a cow — this supports option C. The frames from Chunk 3 show birds in a field, not kids. Since the question is about kids in a paddy field, and Chunk 3 contains no kids or relevant activity, it does not support any option. I will describe what is visible: birds flying and standing in grass. No kids are present. I will not assign an answer since nothing relevant is visible.

Observation: "Multiple black birds are flying and standing in a grassy field, with one white bird visible among them." -3.21

Final answer: C ✓

TimeSearch-R

Call 1. To answer the question about what the two kids are doing in the paddy field, I need to identify the relevant segment in the video. Based on the provided frames and the question, I will search for scenes involving children in a paddy field. I will use the seek_video_frames tool to locate the segment where the two kids are present in a paddy field.

Query: "two kids in a paddy field" *Range:* 0.0–458.6 s, 8 frames.

Call 2. The frames provided show various scenes including a hippopotamus, birds flying over a marsh, fish swimming underwater, and a bird feeding its chick. However, the frames do not clearly depict two kids in a paddy field. I need to search for the specific segment where the two kids are present in a paddy field. I will use the seek_video_frames tool to locate this segment.

Query: "two kids in a paddy field" *Range:* 0.0–458.6 s, 8 frames.

Closing reasoning. The frames clearly show two kids in a paddy field. They are walking through the field, which is consistent with the question. Based on these frames, I can conclude that the two kids are walking in the paddy field.

Final answer: A ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The image shows two birds, one of which appears to be feeding a chick. The chick is nestled on the ground, and the bird with the orange beak is holding a small fish in its beak. This suggests that the bird is likely feeding the chick.

Wait, let's break it down further. The bird's posture and the way it's holding the fish indicate that it's in the process of feeding the chick. The chick is not moving much, which is typical behavior when a chick is being fed. It's staying still, possibly to avoid startling the bird or to receive the food.

Hmm, I need to consider if there's any other possibility. Could the bird be playing with the chick? No, that doesn't seem likely given the context. The bird is holding a fish, which is a food item, and the chick is in a position that suggests it's being fed.

Oh, I see. The two kids in the paddy field are actually the birds, and they are engaged in the act of feeding. The chick is the offspring of the bird, and the bird is providing it with food. This is a common behavior in many bird species where the parents feed their young.

Let's verify this one last time. The image clearly shows a bird feeding a chick, and there's no indication of any other activity. The focus is on the feeding process, and the chick is the recipient of the food.

So, the two kids in the paddy field are the birds, and they are feeding the chick.

Final answer: The two kids in the paddy field are the birds feeding the chick.

Final answer: A ×

Example 7 (`z6HTO2SOxUc_0`). LVB, 4m37s, 7 turns, $\Delta H = +0.64$ bits.
Q: The table holds a black surface with a plate, a multicolored dotted face mask, a black-and-green cucumber model, and a yellow 'Blondies' label. What other objects appear in the video?
Options: (A) Piano (B) Mobile phone (C) Television (D) Leaf **GT:** D
Outcomes: BeliefSearch **D** ✓ · TimeSearch-R **A** × · Video-R1 **A** ×

BeliefSearch

Hypothesis h: A leaf is visible.

T1, c4. In Chunk 4, I see a close-up of a bowl with a powdery substance being mixed, and later a woman in an apron handling a bowl near a stove. There is no visible leaf, phone, TV, or piano. The frames show kitchen activity with bowls, powder, and candy, but no object matching the options. Since the question asks for other objects besides the described ones, and none of the options are visible, I cannot assign an answer yet. I'll continue searching.

Observation: "A person mixes powder in a bowl near a stove, with candy and a Jello box visible nearby." -2.16

T2, c1. The current chunk (Chunk 1) shows people in a kitchen setting with Jello boxes on the table, but no black surface, plate, mask, cucumber model, or yellow label with 'Blondies' is visible. The frames focus on people's faces and the Jello boxes, with no evidence of the described objects. Since the question asks for other objects that appear, and none of the options (piano, phone, TV, leaf) are visible here, I cannot assign an answer. I'll describe what is present: people and Jello boxes. The belief score dropped after searching Chunk 4, so I'll prioritize Chunk 8 next, as it has the highest belief and shows a man eating, which might reveal more context.

Observation: "People are seated at a table with Jello boxes, smiling and interacting, with no black surface, plate, mask, cucumber model, or yellow label visible." -2.55

T3, c8. The frames from Chunk 8 show a man eating from a bowl, with a Jello box and candy visible on the table. There is no black surface, plate with a mask, cucumber model, or yellow label. The scene is a kitchen with white cabinets and a stove. No leaf, phone, TV, or piano is visible. The most relevant object from the options is a leaf, as the table has candy and a Jello box, which might imply autumn decor, but no leaf is clearly present. Since the question asks for other objects and none of the options are visible, I must describe what is present. The frames show people eating and a Jello box, which is not an option. I'll describe the visible objects without mentioning absence. The hypothesis was "A leaf is visible," but no leaf is seen. I'll output an observation for Chunk 8, describing what is present, and since no option is supported, I'll omit "answer." I'll search Chunk 7 next, as it has the highest belief and shows candy and Jello, which might relate to the yellow label.

Observation: "A man eats from a bowl at a kitchen counter with a Jello box and candy visible on the table, and white cabinets in the background." -2.21

T4, c6. The frames from Chunk 6 show a man mixing ingredients in a glass bowl on a white kitchen counter next to a stove. A Jello box is visible on the counter, and the man is wearing an apron. There is no black surface, plate, face mask, cucumber model, or yellow label with "Blondies" visible. The scene is focused on food preparation, with no objects matching the question's description. Since the question asks for other objects that appear, and none of the options (piano, phone, TV, leaf) are visible, I must describe what is present. The most relevant object is the Jello box, but it's not an option. I'll describe the mixing action and Jello box as present. No option is supported, so I omit "answer".

Observation: "A man mixes ingredients in a glass bowl on a white counter next to a stove, with a Jello box visible nearby." -2.08

T5, c3. The question asks for other objects besides the described items. In Chunk 3, I see a man in a kitchen wearing an apron, mixing ingredients in a bowl on a white counter. A Jello box is visible on the counter, and there are autumn-themed decorations including leaves and pumpkins. The yellow label with "Blondies" is visible in one frame, and the black surface is the counter. The cucumber model is not visible, but the leaves are clearly present. Since "leaf" is an option and leaves are visible, this supports option D. I'll describe the leaves and the Jello box as relevant objects.

Observation: "A man mixes ingredients in a bowl on a white counter with a Jello box and autumn leaves visible nearby." $+1.46$, ans $\rightarrow$ D

T6, c2. The search history shows that Chunk 3 (blurry kitchen) received a +1 score because it mentioned "autumn leaves visible nearby," which aligns with option (D) Leaf. The current chunk (Chunk 2) shows a man in a kitchen with a Jello box and a bowl of eggs, but no black surface, plate, mask, cucumber model, or yellow label. However, the man is wearing an apron with leaf patterns, and the background includes a plant and autumn-themed items. Since the question asks for other objects that appear, and leaves are visible in the background (e.g., in Chunk 3 and possibly here), I can infer that a leaf is present. The hypothesis is "A leaf is visible," and the frames show leaf-patterned apron and plant, which supports (D). I'll output the observation with answer D.

Observation: "A man in a kitchen wears a leaf-patterned apron, and a potted plant with green leaves is visible on the counter." $+1.32$, ans $\rightarrow$ D

Final answer: D ✓

TimeSearch-R

Call 1. To answer the question, I need to identify the objects mentioned in the query and then search for other objects that appear in the video. The objects in the query are a black surface, a plate, a multicolored dotted face mask, a black and green cucumber model, and a yellow label with the word 'Blondies'. I will search for these objects in the video and then look for other objects that appear alongside them.

Query: "black surface on the table with a plate on it" *Range:* 0.0–277.2 s, 8 frames.

Call 2. From the provided frames, I can see that the black surface on the table with a plate on it contains a multicolored dotted face mask, a black and green cucumber model, and a yellow label with the word 'Blondies'. I will now search for other objects that appear in the video alongside these items.

Query: "multicolored dotted face mask" *Range:* 0.0–277.2 s, 8 frames.

Call 3. From the provided frames, I can see that the black surface on the table with a plate on it contains a multicolored dotted face mask, a black and green cucumber model, and a yellow label with the word 'Blondies'. The frames also show a man in a plaid shirt and a woman in a striped shirt. Based on these frames, I will now search for other objects that appear in the video alongside these items.

Query: "black and green cucumber model" *Range:* 0.0–277.2 s, 8 frames.

Final answer: A ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The question is asking about objects that appear in the video, but it doesn't specify which objects are already mentioned. So, I need to carefully consider the context and the objects that are likely to be present in the video.

First, let's break it down. The video shows a black surface with a plate, a multicolored dotted face mask, a black and green cucumber model, and a yellow label with the word 'Blondies'. These objects are already mentioned, so they don't need to be counted again.

Now, let's think about what other objects might appear in the video. The question doesn't provide a list of objects, so I need to rely on common kitchen items and typical cooking scenes.

Hmm, in a cooking video, it's common to see various kitchen tools and ingredients. For example, there might be a knife, a spoon, a bowl, or some ingredients like flour, sugar, or eggs. These are not explicitly mentioned in the question, but they are likely to be present.

Wait, there's also the possibility of a cooking pot or a pan, which are common in kitchen scenes. These objects are not mentioned in the question, but they are likely to be present.

Oh, I see. The question is asking for objects that are not explicitly mentioned but are typical in a cooking video. So, the objects that appear in the video but are not mentioned in the question are likely to be kitchen tools and ingredients.

Let's verify this. The question doesn't provide a list of objects, so the objects that are not mentioned but are typical in a cooking video are the most likely answer.

Therefore, the answer should be the objects that are not explicitly mentioned but are typical in a cooking video.

Final answer: D

Let's break it down one more time to ensure. The objects that are not mentioned but are typical in a cooking video are the most likely answer.

Final answer: A ×

F.4 Failure cases

Two examples where the search runs to convergence but lands on the wrong answer. Both baselines also fail. The two failures expose distinct mechanism-level limitations.

Example 8 (`videomme_512-1`). VME, 7m33s, 6 turns, $\Delta H = +0.84$ bits.
Q: What do the judges do after the little girl finishes the performance?
Options: (A) They stand up and applaud for her (B) They sit on the seats and applaud for her (C) They blow the little kisses (D) None of the above **GT:** A
Outcomes: BeliefSearch **B** × · TimeSearch-R **B** × · Video-R1 **B** ×

BeliefSearch

Hypothesis h: judges are applauding.

T1, c7. The frames from Chunk 7 show the girl performing on stage with the judges seated at a table in front of her. They are not standing or blowing kisses, and clapping is not explicitly visible, but their seated posture and orientation toward the girl suggest they are reacting to the performance. This is most consistent with option B (sitting and applauding), although no clapping motion is yet captured.

Observation: "The judges are seated at a table in front of the girl on stage, facing her and appearing to react to her performance." -2.70, ans $\rightarrow$ B

T2, c1. Chunk 1 shows a little girl in a green top performing on stage with other girls and adults watching. The judges are not clearly visible in these frames, so this chunk provides no evidence for any option. I will continue searching.

Observation: "A little girl in a green top performs on stage while other girls and adults watch, and no judges are visible in these frames." -3.05

T3, c10. The frames from Chunk 10 show the girl finishing her performance and being embraced by a woman in a black jacket, likely a judge or host. The woman is seated, and the girl stands on stage. No standing or applauding is visible from the judges, and no one is blowing kisses. The judges are seated and reacting, which still aligns with option B.

Observation: "A woman in a black jacket is seated and hugging the little girl on stage after her performance." -3.49, ans $\rightarrow$ B

T4, c4. Chunk 4 shows the girl performing while a judge in a pink top is clapping while seated, and another judge in a black jacket waves at her. The judges are not standing, and no one is blowing kisses. Clapping while seated directly matches option B.

Observation: "A judge in a pink top is seated and clapping while watching the girl perform, and another judge in a black jacket waves at the girl while seated." $+1.56$, ans $\rightarrow$ B

T5, c5. The current belief is highest for Chunk 4, which already showed judges seated and clapping. Chunk 5 again shows a judge in a pink top seated and clapping while the girl performs, the same pattern as the previous turn. This continues to support option B.

Observation: "A judge in a pink top is seated and clapping while watching the girl perform on stage." $+1.27$, ans $\rightarrow$ B

T6, c3. Chunk 3 shows the girl dancing on stage with a man in a black jacket pointing at her, and a judge in a pink top still seated and clapping. The pattern of seated, clapping judges holds across every chunk that shows them. I have enough evidence to answer.

Final answer: B $\times$

TimeSearch-R

Call 1. To answer this question, I need to identify the part of the video where the little girl finishes her performance and the judges' reactions. I will search for the segment where the girl finishes her dance and the judges' reactions follow.

Query: "little girl finishes performance" *Range:* 180.0–453.6 s, 8 frames.

Closing reasoning. The frames clearly show the little girl finishing her performance on stage. The judges are seen clapping and reacting positively to her performance. Based on these frames, I can conclude that the judges stand up and applaud for her.

Final answer: B $\times$

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The question is asking what the judges do after the little girl finishes her performance. Now, let's break it down step by step.

First, I need to consider the context of the image. The little girl has just finished her performance, and the judges are seated in front of her. They are likely waiting for her to finish before reacting.

Now, let's look at the options: A. They stand up and applaud for her. - This seems plausible because it's a common reaction after a performance. B. They sit on the seats and applaud for her. - This is also possible, but it's less likely because standing up is a more enthusiastic and immediate reaction. C. They blow the little kisses. - This is not a typical reaction after a performance, so it's unlikely. D. None of the above. - This is not a valid option since we have three other options to choose from.

Hmm, it seems that the most likely reaction is that the judges stand up and applaud for her. This is a common and expected behavior in such situations.

Oh, I see. The judges are likely to stand up and applaud for her after she finishes her performance. This is the most logical and common reaction.

Let's verify this one last time. The judges are seated, and the little girl has just finished her performance. It's natural for them to stand up and applaud as a sign of appreciation.

After this reflection, I am confident that the correct answer is A.

Let's break it down one last time to ensure. The judges are seated, the little girl has finished her performance, and the most common and expected reaction is to stand up and applaud.

Final answer: A

Final answer: **B** ×

Example 9 (`videomme_397-3`). VME, 10m28s, 7 turns, $\Delta H = +1.09$ bits.
Q: What story does the video tell?
Options: (A) Who is the ancestor of birds (B) How fossil comes about (C) We should visit the dinosaur museum (D) Dinosaurs went where **GT:** D
Outcomes: BeliefSearch **A** × · TimeSearch-R **B** × · Video-R1 **B** ×

BeliefSearch

Hypothesis h: Dinosaurs and birds are shown together.

T1, c9. The frames in Chunk 9 show a stylized poster of a dinosaur with a person, labeled "THE VALLEY OF GWANGI," and a small dinosaur in a natural setting with flowers. The poster suggests a narrative or exhibit, and the presence of both dinosaurs and birds in earlier chunks implies a focus on their relationship. The most relevant visual evidence for the question is the coexistence of dinosaurs and birds, which supports option A. I will describe the visible content without referencing the option directly.

Observation: "A poster depicts a large dinosaur with a person, and a small dinosaur is shown in a grassy area with flowers." -2.46, ans $\rightarrow$ A

T2, c1. The frames from Chunk 1 show a stylized animated sequence with a large dinosaur silhouette against a night sky, a bird perched on a branch, and a dinosaur near a large egg-like object. The presence of both dinosaurs and birds, along with the egg, suggests a narrative about evolution or ancestry. The bird is shown in a natural setting, and the dinosaur appears in various poses, including near the egg, which could imply a connection between them. The most relevant option is (A) because the visual evidence points to a relationship between dinosaurs and birds, possibly indicating that dinosaurs are the ancestors of birds. The description should focus on what is visible: the bird and dinosaur together, and the egg, without mentioning locations or option details.

Observation: "A bird is perched on a branch near a large egg-like object, and a dinosaur is shown in silhouette against a night sky." $+1.68$, ans $\rightarrow$ A

T3, c2. The current chunk shows a large dinosaur with a textured hide and prominent horns, standing near a body of water with a smaller dinosaur visible in the background. The setting includes a sunset sky and sparse vegetation, suggesting a natural prehistoric environment. This visual content supports the idea that the video is telling a story about dinosaurs in their natural habitat, which aligns with option D, as it focuses on where dinosaurs went rather than their ancestry, fossils, or museum visits. The presence of multiple dinosaurs in a natural setting reinforces the narrative of movement or presence in specific locations.

Observation: "A large dinosaur with horns stands near water with a smaller dinosaur in the background under a sunset sky." -1.65, ans $\rightarrow$ D

T4, c5. The current chunk (5) shows a large dinosaur in a natural setting with a sunset and water, and a smaller dinosaur in the background. The presence of both large and small dinosaurs in a prehistoric environment suggests a narrative about dinosaur evolution or movement, but no direct evidence of birds or fossils is visible here. The frames do not show a bird or a fossil, so this chunk does not support option A or B. The setting is outdoors, not a museum, so option C is not supported. The dinosaurs are shown in a landscape, which could relate to option D, but the frames do not show them moving to a specific location. Since the question is about the story told, and the video includes both dinosaurs and birds in other chunks, the most consistent narrative is about the evolutionary link between them. However, this chunk alone does not provide enough evidence to confirm any option. I will describe what is present: a large dinosaur and a smaller dinosaur in a natural setting with a sunset.

Observation: "A large dinosaur and a smaller dinosaur are visible in a natural landscape with a sunset and water in the background." -2.17, ans $\rightarrow$ D

T5, c10. The current chunk (10) shows a large dinosaur in a natural setting with a person holding a notebook, and a screen displaying a list of names, suggesting a documentary or educational context. The presence of both the dinosaur and the person implies a narrative about observation or study, which aligns with option (C) "We should visit the dinosaur museum" if the setting is a museum exhibit. However, the frames also show the dinosaur in a natural environment, which could support (D). The key is to focus on what is visible: a person interacting with a dinosaur display and a list of names, which suggests a museum visit. Since the question is about the story told, and the visual evidence points to a museum context with a dinosaur, (C) is the most supported option. I will describe the visible elements without negation and assign the answer.

Observation: "A large dinosaur is displayed in a museum-like setting with a person holding a notebook and a screen showing names, suggesting an educational or exhibit context." -3.44, ans $\rightarrow$ C

T6, c6. The current belief is highest for Chunk 1 (0.51), which showed a bird on a branch and a dinosaur in silhouette, supporting option A. Chunk 6 frames show a large dinosaur in a natural setting with a sunset, but no bird or direct evolutionary link. The frames depict a T-Rex-like dinosaur with detailed skin texture and a sunset background, suggesting a focus on dinosaur presence rather than ancestry or fossils. Since the question is about the story told, and the video includes both dinosaurs and birds in some frames, the most consistent evidence points to the evolutionary relationship between them. I will now search Chunk 7, which is labeled "Dinosaur with bird" and has a belief of 0.08, to confirm if it shows a direct link.

Observation: "A large dinosaur with detailed skin texture is shown in a natural landscape with a sunset, and a smaller dinosaur is visible in the background." -2.53, ans $\rightarrow$ D

Final answer: A ×

TimeSearch-R

Call 1. To determine the story the video tells, I need to analyze the content across the frames. The video features various scenes involving dinosaurs in different environments, such as a night sky, a lake, and a museum. The frames suggest a narrative that spans from the life of dinosaurs to their fossilization and display in a museum. I will use the seek_video_frames tool to identify key scenes that help in understanding the story.

Query: "dinosaurs in a museum" *Range:* 500.0–628.3 s, 8 frames.

Call 2. The frames clearly depict the process of dinosaur fossils being discovered and displayed in a museum. The frames show dinosaur skeletons, museum displays, and even a dinosaur skeleton being reconstructed. This suggests that the video is telling the story of how fossils come about. I will now analyze the frames to confirm this story.

Query: "fossilization process" *Range:* 0.0–500.0 s, 8 frames.

Closing reasoning. The frames clearly illustrate the life cycle of dinosaurs, from their habitat to their fossilization process. The frames show dinosaurs in their natural environment, a bird laying eggs, and the fossilization of a dinosaur. This sequence of frames supports the story of how fossils come about. I will now provide the answer based on these frames.

Final answer: B ×

Video-R1

Full reasoning chain (verbatim).

Let me think about this for a moment. The video seems to be a collection of images that depict various dinosaurs and fossils, which suggests that it's telling a story about dinosaurs.

Wait, let's break it down. The images show dinosaurs in different settings, like a mountain, a lake, and a museum. The presence of fossils and skeletons also indicates that the video is likely about the history and evolution of dinosaurs.

Hmm, option A, "Who is the ancestor of birds," is intriguing. The video does show a dinosaur with feathers, which could be related to the evolution of birds. However, the video doesn't seem to focus solely on this aspect.

Option B, "How fossil comes about," is also plausible. The video shows fossils and skeletons, which could be used to explain the process of fossilization. However, the video doesn't seem to delve deeply into this topic either.

Option C, "We should visit the dinosaur museum," is not a strong contender. While the video does show a museum, it doesn't seem to be the main focus of the story.

Oh, I see. Option D, "Dinosaurs went where," is the most fitting. The video shows dinosaurs in different settings, which could be interpreted as their migration or movement across the landscape. This option aligns well with the visual content of the video.

Let's verify this. The video does not provide a detailed explanation of the fossilization process or the evolution of birds. Instead, it focuses on the presence and movement of dinosaurs in different environments. Therefore, the most accurate answer is D.

Final answer: D

Let's double-check. The video does not provide a comprehensive answer to any of the other options, but it does suggest that the story is about the movement and presence of dinosaurs.

Final answer: B ×

What the failures reveal. The two failure cases share one underlying mechanism: the hypothesis h chosen at the preview turn commits the search to a single interpretation of the question, and once one observation matches that interpretation, the asymmetric kernel cannot redirect mass to a different option even when later observations would support it.

In Example 8, the correct answer (A, stand up and applaud) requires evidence specifically about standing. The hypothesis the model emits, "judges are applauding," is a property shared by both A and B and does not distinguish them. Once T4 finds judges seated and clapping, the observation entails the hypothesis at $+1.56$, the belief locks on B, and a second positive observation in T5 reinforces that mass without ever prompting a search for the standing-up moment. The scorer was not wrong on its own terms (the seated-clapping observation genuinely entailed the hypothesis), but the hypothesis was too coarse to pick out the option the question was actually asking about.

In Example 9, the correct answer (D, where dinosaurs went) is a narrative arc that no single chunk can encode. The hypothesis, "dinosaurs and birds are shown together," aligns with option A (ancestor of birds) and produces a $+1.68$ match on chunk 1 at T2. The trajectory then oscillates among A, D, and C as different chunks suggest different threads, but none produce positive evidence strong enough to tilt the belief away from A. The belief does converge ($\Delta H = +1.09$ bits), just onto the wrong option.

Both failures map onto the limitations identified in §6: a single declarative hypothesis cannot represent every option a multiple-choice question puts in play, and once positive evidence has been

locked in, the asymmetric kernel that propagates negative evidence locally cannot redirect the belief globally. A learned scorer trained jointly with the controller, or a multi-hypothesis variant that scores observations against several candidate interpretations in parallel, would address both cases within the same Bayesian framework.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We state our main claims in the abstract and §1, and they match the contributions of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in §6 and in the failure-case analysis in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not contain formal theorems or proofs. The probabilistic derivations in §4 are presented as method specification rather than theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the method in §4 and disclose all hyperparameters, training stages, and code-level choices in Appendix A and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used in our experiments are publicly available, and we will release our code and model weights upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We specify training and test settings in §5 and provide the full hyperparameter and configuration details in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Following common practice for evaluation on VideoMME, LongVideoBench, and MLVU, we report single-run numbers without error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type and scale used for training and inference in §5 and provide additional compute details in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics. We use only publicly available datasets and do not involve human subjects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work has dual-use potential. Positive applications include automatic content moderation, video accessibility (captioning and summarization), evidence search, and the transfer of the decoupled belief framework to long-document QA and embodied search. Dual-use risks include large-scale surveillance of recorded video and automated profiling from CCTV or social-media footage. We did not introduce a new dataset or benchmark beyond those already public, and the system's explicit-belief design makes its

segment choices and stopping decisions auditable, which can support oversight and post-hoc review without eliminating the underlying misuse risks.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our model is a fine-tuned video question answering system and does not pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and pretrained models we use are publicly released and cited in §5.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The method and training procedure are documented in §4 and Appendix A, and the released code and model weights will include accompanying documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve research with human subjects, so IRB review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our method uses VLMs (Qwen2.5-VL as the controllable model and Qwen3-VL-235B-A22B as a teacher for trace distillation) as core components, and we document their usage throughout §4 and §5.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.