WHAT HAPPENED VS. WHAT CAN BE KNOWN: ONTIC–EPISTEMIC VIDEO REVISIT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video-language models are increasingly capable of recognizing temporal state changes, yet they often conflate what happened in the world with what can be concluded from the visual evidence they actually observed. We study this distinction through ontic–epistemic reasoning across video revisits, where a model must compare entity states across temporally separated visits, recognize a physical transition, and respond appropriately when visit-specific evidence is incomplete. We construct structured *Revisit Packs* from paired observations of the same environment. Each pack contains controlled variants that remove or perturb visit-specific evidence while preserving the underlying physical case. We first extend Behavior Pack Optimization (BPO), originally designed around counterfactual views of a single video-question instance, to these paired revisits so that cross-visit transition behavior can be optimized jointly. Although BPO provides anchor-relative updates across the pack, it does not explicitly prioritize the intervention relations that remain unsatisfied. We therefore introduce *Intervention-Structured Bottleneck Behavior Pack Optimization* (BBPO), which uses detached pack-local relation-success scores to redistribute BPO advantage toward poorly satisfied revisit relations while preserving the mean optimization scale of the pack. Experiments show that BBPO raises mean all-view pack correctness from 63.6% to 66.3%, including improvements from 54.0% to 57.5% on Revisit Packs and from 68.4% to 70.7% on Visit Packs. Extensive results suggest that reliable reasoning across video revisits benefits not only from jointly optimizing evidence interventions, but also from allocating post-training credit according to the behavioral relation that most limits all-view revisit correctness.

1 INTRODUCTION

Interaction between multimodal agents and changing environments requires video understanding beyond isolated observations. Recent video multimodal large language models (MLLMs) have introduced diverse memory mechanisms that substantially improve the ability to store and retrieve temporally distant evidence Song et al. (2024); He et al. (2024); Long et al. (2026). However, access to past observations does not by itself guarantee correct reasoning across revisits. When an environment is revisited, a model must associate entities across temporally separated observations, retain their previously observed states, and reason about how those states may have changed. Such reasoning is fundamental to persistent visual understanding: recognizing the current state alone is insufficient if the model cannot relate it to what was observed during an earlier visit.

This challenge becomes increasingly important as video understanding shifts from retrieving past events toward maintaining persistent representations of a changing world. For example, M^3-Verse evaluates MLLMs using paired egocentric observations of the same indoor environment before and after state changes, explicitly probing whether models can recognize transformations across revisits Wei et al. (2025). Recent benchmarks further establish persistent state reasoning as an important capability, requiring models to retain entity states and reason about how they evolve over time Wang et al. (2026); Yu et al. (2026). These works primarily focus on recovering or reasoning about the state trajectory of the underlying world. Our work introduces a complementary question: when the evidence available from one or both visits is incomplete, can the model distinguish the transition that physically occurred from the transition that is actually justified by what it observed? Recent studies have shown that MLLMs often continue to answer even when the visual evidence required by the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

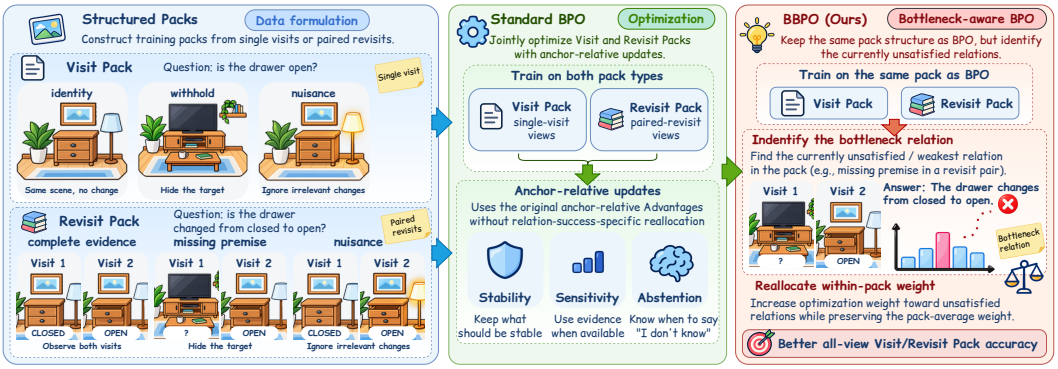

Figure 1: The pack design instantiates *ontic* as: what physical state or transition actually occurred, and *epistemic* as: what the available visual evidence is sufficient to justify. Standard BPO is adapted to revisit pairs, and the proposed BBPO aims to enhance all-view correctness.

question is missing Whitehead et al. (2022); Madhusudhan et al. (2026); pramono et al. (2026). Reliable revisit reasoning should respond appropriately to evidence availability: commit to a transition when the relevant states are observed, recognize when a required visit-specific premise is missing, and remain stable when irrelevant visual content is perturbed.

To address that, we instantiate the ontic–epistemic video revisit reasoning. The *ontic* component describes the physical states and transition that actually occurred, whereas the *epistemic* component concerns what the model is justified in concluding from the evidence available in the visits. Crucially, the physical history remains fixed while the model's epistemic access can be changed through controlled interventions on the observations. The intervention-induced behavioral relations provide a concrete way to evaluate whether a model distinguishes changes in the world from changes in what can be known about that world. The overview of our work is shown in Fig. 1.

Behavior Pack Optimization (BPO) provides a natural mechanism for learning evidence-sensitive behavior Kang et al. (2026). BPO begins from a single video-question instance, constructs question-type-conditioned counterfactual views of that instance, and jointly evaluates whether the model remains stable under irrelevant interventions, changes its answer when diagnostic evidence is altered, or abstains when the required evidence is removed. Rather than probing a single video-question instance, we reason over paired video revisits of the same environment. We therefore adapt the behavior-pack formulation to revisit pairs, organizing evidence-complete, evidence-withheld, and nuisance-modified observations around the same underlying physical transition. In practice, some relations may already be handled correctly while one remaining failure determines the correctness of the entire pack. Standard anchor-relative BPO assigns each non-anchor view its reward-relative advantage, but it does not explicitly redistribute optimization pressure according to which intervention relation currently constitutes this pack-level bottleneck.

Hence, we introduce Intervention-Structured Bottleneck Behavior Pack Optimization (BBPO), which retains the original BPO view loss and anchor-relative formulation but augments its credit allocation with a detached, pack-local measure of intervention success. Unsatisfied relations receive greater optimization weight, while a uniform component preserves learning pressure on the remaining views. The weights are normalized within each pack, such that BBPO reallocates rather than globally increases optimization pressure. Across three matched seeds, BBPO improves mean all-view accuracy from from 63.6% to 66.3%, including gains from 54.0% to 57.5% on Revisit Packs and from 68.4% to 70.7% on Visit Packs. These results indicate that the benefit comes from improving joint behavioral correctness across revisit interventions rather than from uniformly improving every uncertainty-related diagnostic. Overall, our study reframes video memory under revisitation from identifying a single physical history to reasoning about the boundary between what happened and what can be known from what was observed. Our contributions are threefold:

- We formulate ontic–epistemic video revisit reasoning, distinguishing the physical transition that occurred from the transition justified by the evidence available across two separated visits.

- We extend behavior-pack post-training from single-video counterfactual QA to paired video revisits, constructing intervention packs that probe transition recognition, missing-evidence behavior, and invariance to irrelevant changes.

- We propose Intervention-Structured Bottleneck Behavior Pack Optimization (BBPO), which retains BPO's anchor-relative objective while reallocating within-pack optimization pressure toward poorly satisfied revisit relations, thereby improving all-view recognition.

2 Related Work

Long-Video Memory. MLLMs Zhang et al. (2026); Chen et al. (2026); Liu et al. (2026) have substantially extended video understanding beyond short clips, but processing long visual streams remains challenging because of growing token cost, limited context capacity, and the difficulty of preserving temporally distant information. A prominent direction addresses these limitations through explicit memory mechanisms He et al. (2024); Song et al. (2024). Recent works organize long-term information into structured episodic, semantic, and visual memories that can be selectively retrieved during reasoning Long et al. (2026); Yeo et al. (2026). These methods primarily address *what information should be retained and retrieved* so that evidence remains accessible over long temporal horizons. Our work is complementary: even a perfect memory of the available observations cannot recover a state that was never observed. We instead study how a video model should reason when its persistent visual memory is itself epistemically incomplete.

Persistent State Memory and Revisit Reasoning. Beyond retaining isolated events, it is important to examine whether MLLMs can maintain a coherent representation of a world whose entities change over time Wei et al. (2025); Yu et al. (2026); Wang et al. (2026). Particularly, M^3-Verse explicitly evaluates reasoning across paired egocentric observations of the same indoor environment before and after state changes Wei et al. (2025). These studies move video understanding toward *persistent state memory*: remembering what state an entity occupied and how that state evolved. However, recovering the physical state trajectory and determining what can be inferred about that trajectory from the available observations are distinct problems. A physical transition may be well defined in the underlying world while remaining unsupported when evidence from one of the visits is unavailable. Our work focuses on this latter issue, asking not only *what happened across two visits*, but also *whether the available evidence is sufficient to support that conclusion*.

Reasoning with Missing Visual Evidence. Recent work studies whether MLLMs recognize when visual evidence is insufficient to support an answer Kirichenko et al. (2026); Zhu et al. (2026); Zhou et al. (2026); pramono et al. (2026). Behavior Pack Optimization (BPO) goes beyond single-view supervision by constructing, for each video-question instance, a compact set of question-type-conditioned counterfactual views and optimizing their outputs jointly Kang et al. (2026). Its behavior contracts encourage stability under label-preserving interventions, sensitivity to answer-changing evidence manipulations, and honest abstention when diagnostic evidence is removed. BPO therefore establishes an important framework for evidence-sensitive video post-training. Our work extends BPO to paired video revisits, where the relevant behavior depends jointly on observations from two temporally separated visits. Rather than only testing whether one video's answer remains supported under manipulation, we construct revisit-centered intervention packs that probe cross-visit transition recognition, missing visit-specific evidence, and invariance to irrelevant changes.

Ontic-Epistemic Reasoning. The separation between changes in the world and changes in an agent's knowledge has a long history. The dynamics of knowledge distinguishes *ontic* actions that alter the underlying world from *epistemic* changes that alter what an agent knows about that world Grossi et al. (2017). More recent work studies epistemic reasoning under partial information, including what can be inferred from incomplete observations or another agent's access to information Annese et al. (2025); Gabay et al. (2026). We instantiate this distinction in paired video revisits. The ontic variable is the physical state transition between two visits, while the epistemic condition is determined by which visit-specific evidence is accessible to the model. By holding the physical transition fixed while removing or perturbing evidence from the visits, we can test whether model behavior responds appropriately to changes in epistemic access rather than merely reproducing the hidden physical trajectory.

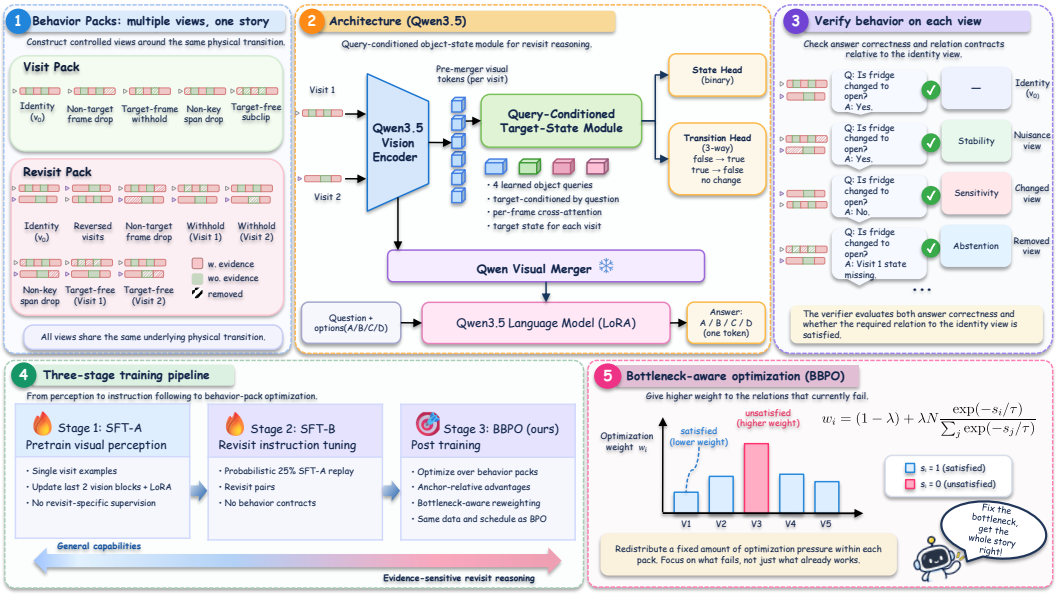

Figure 2: Framework of our method. (1) Behavior Pack construction (Sec. 3.1). (2) Architecture with backbone Qwen3.5 (Sec. B). (3) Behavioral contract, (4) training pipeline, and (5) BBPO are explicitly described in Sec. 3.2 – 3.4, respectively.

3 Ontic–Epistemic Video Revisit Reasoning

A paired video revisit is $(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)})$, where $\boldsymbol{V}^{(1)}$ and $\boldsymbol{V}^{(2)}$ are temporally separated observations of the same physical environment. Given a question q about an entity and its state across the two visits, let $z^{(1)}$ and $z^{(2)}$ denote the corresponding physical states and $t^\star = z^{(1)} \to z^{(2)}$ the physical transition that actually occurred. Our framework is illustrated in Fig. 2, while architectural details are provided in Appendix Sec. B.

We use *ontic* to denote reasoning about the physical states and transitions that occurred across visits, and *epistemic* to denote reasoning about whether those states or transitions are justified by the visual evidence available to the model. Evidence-complete Visit and Revisit Packs primarily evaluate ontic recognition, whereas evidence-incomplete Revisit Packs evaluate whether the model appropriately adjusts its conclusions when the required visit-specific evidence is unavailable.

3.1 Visit Packs and Revisit Packs

We define following structured behavior packs. A *Visit Pack* contains controlled views associated with an individual visit and probes local state recognition and evidence sensitivity without requiring cross-visit transition reasoning, which contains 5 views of a single visit: identity, non-target-frame drop, non-key-span drop, target-frame withholding, and a natural target-free subclip. A *Revisit Pack* is constructed around a paired revisit $(\boldsymbol{V}^{(1)}, \boldsymbol{V}^{(2)})$ and evaluates reasoning about the state transition across the two visits, which contains 8 views of the paired revisit: identity, reversed visits, non-target-frame drop, Visit-1 and Visit-2 target-frame withholding, non-key-span drop, and Visit-1 and Visit-2 target-free subclips. In contrast to BPO, we use natural evidence withholding instead of direct evidence masking. Detail is described in Appendix Sec. C.

For a pack $\mathcal{B}$, the evidence-complete identity view v_0 serves as the anchor, while $v_1, \ldots, v_N$ are intervention views. Each view is associated with a target response $y_i^\star$, behavioral contract c_i, and intervention metadata ξ_i, denoted by

$$\omega_i = (v_i, y_i^\star, c_i, \xi_i). \tag{1}$$

And the structure of a pack is represented by $E(\mathcal{B}) = \{(v_0, v_i) \mid i = 1, \ldots, N\}$.

3.2 Behavior Pack Optimization for Paired Revisits

Behavior Pack Optimization (BPO) Kang et al. (2026) constructs counterfactual views of an original video-question instance and jointly optimizes their outputs using view-level rewards and cross-view behavioral contracts. Its contracts encourage stability under label-preserving interventions, sensitivity to answer-changing interventions, and abstention when diagnostic evidence is removed. BPO further uses the original, unintervened view as a prompt-specific anchor for the non-anchor updates.

We adapt this framework from counterfactual views of a single video-question instance to paired video revisits. The anchor v_0 contains the complete revisit evidence, while each v_i alters the availability or relevance of evidence from one or both visits. For each view, the policy generates $y_i \sim \pi_\theta(\cdot \mid v_i, q)$, and a verifier assigns a reward $R_i = R(v_i, y_i; y_0, c_i)$, which incorporates both view-level correctness and satisfaction of the required behavioral contract relative to the identity view. Therefore, the BPO objective is

$$\mathcal{L}_{\text{BPO}} = -\frac{1}{N}\sum_{i=1}^{N} \min\left[\rho_i(\theta)A_i^{\text{BPO}}, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\, A_i^{\text{BPO}}\right] + \beta\widehat{\mathcal{L}}_{\text{KL}}, \tag{2}$$

where $A_i^{\text{BPO}} = R_i - R_0, i \in \{1, \ldots, N\}$ denotes the anchor-relative advantage for each view, $\rho_i(\theta) = \frac{\pi_\theta(y_i|v_i,q)}{\pi_{\text{old}}(y_i|v_i,q)}$ denotes the policy importance ratio, and $\widehat{\mathcal{L}}_{\text{KL}}$ is a sampled log-ratio regularizer that regularizes the policy toward a reference policy. RLHF-style PPO typically regularizes the learned policy toward a fixed SFT reference policy via a KL penalty Ouyang et al. (2022). In contrast, we use a partially adaptive reference whose visual and object-state representations are shared with the current policy, while the task-specific language LoRA adapter is disabled and the reference score is stop-gradient. More detail will be described in Sec. 3.3.

Applying BPO to paired revisits already encourages evidence-sensitive cross-view behavior. However, standard BPO does not explicitly condition the allocation of optimization pressure on whether each behavioral relation within the current pack is already satisfied. This becomes important because all-view pack correctness is conjunctive. Let $C_i^{\text{ans}} \in \{0, 1\}$ indicate whether the answer for view v_i is correct and $C_i^{\text{rel}} \in \{0, 1\}$ indicate whether its required relation to the identity view is satisfied. We define

$$C_{\text{pack}}(\mathcal{B}) = C_0^{\text{ans}} \prod_{i=1}^{N} C_i^{\text{ans}} C_i^{\text{rel}}. \tag{3}$$

Hence, $C_{\text{pack}}(\mathcal{B}) = 1$ only when the identity answer and every non-anchor answer–relation pair are correct. For example, a pack with correctness pattern $\{1, 1, 1, 0\}$ attains 75% marginal correctness but remains an incorrect all-view pack. A single unsatisfied relation can determine the correctness of the entire pack, motivating the bottleneck-aware within-pack credit allocation introduced next.

3.3 Intervention-Structured Bottleneck BPO

We introduce *Intervention-Structured Bottleneck Behavior Pack Optimization* (BBPO), which reweights each non-anchor advantage according to whether its intervention relation is satisfied.

Relation success. For each relation $(v_0, v_i) \in E(\mathcal{B})$, we define the detached binary success score

$$s_i = \mathbb{I}[C_0^{\text{ans}} = 1]\, \mathbb{I}[C_i^{\text{ans}} = 1]\, \mathbb{I}[C_i^{\text{rel}} = 1]. \tag{4}$$

Specifically,

$$C_i^{\text{rel}} = \begin{cases} C_i^{\text{stab}}, & c_i = \texttt{stability}, \\ C_i^{\text{sens}}, & c_i = \texttt{sensitivity}, \\ C_i^{\text{abst}}, & c_i = \texttt{abstention}. \end{cases} \tag{5}$$

For `stability`, the intervened prediction should remain consistent with the identity prediction. For `sensitivity`, it should change as required by the intervention. For `abstention`, evidence removal should produce the designated insufficient-evidence behavior rather than retain an unsupported identity conclusion. The score s_i is used only for credit allocation and is detached from the computational graph. All intervention views are described in Appendix Sec. C.

Bottleneck-aware weighting. BBPO assigns greater optimization weight to unsuccessful relations through

$$w_i = (1-\lambda) + \lambda |E(\mathcal{B})| \frac{\exp(-s_i/\tau)}{\sum_{j \in E(\mathcal{B})} \exp(-s_j/\tau)}, \qquad (v_0, v_i) \in E(\mathcal{B}), \tag{6}$$

where $\tau > 0$ controls the sharpness of bottleneck prioritization and $\lambda \in [0,1]$ controls the strength of reallocation. Since $|E(\mathcal{B})| = N$, the weights satisfy $\sum_{i \in E(\mathcal{B})} w_i = N$, and hence their mean is 1 (See proof in Proposition A.1). BBPO therefore redistributes optimization pressure within each pack. rather than globally increasing the contribution of difficult examples. When $\lambda = 0$, $w_i = 1$ for every relation and the weighting reduces to standard BPO.

Proposition 3.1 (Failure-prioritized allocation)**.** *Assume $s_i \in \{0,1\}$, where $s_i = 0$ denotes an unsatisfied relation and $s_i = 1$ a satisfied relation. Let F and S be the numbers of failed and satisfied relations in a pack, respectively, with $F + S = N$. If $F > 0$, $S > 0$, $\lambda > 0$, and $\tau > 0$, then every failed relation receives weight larger than one and every satisfied relation receives weight smaller than one:*

$$w_{\text{fail}} > 1 > w_{\text{succ}}. \tag{7}$$

In particular,

$$w_{\text{fail}} = (1-\lambda) + \lambda N \frac{1}{F + Se^{-1/\tau}}, \qquad w_{\text{succ}} = (1-\lambda) + \lambda N \frac{e^{-1/\tau}}{F + Se^{-1/\tau}}, \tag{8}$$

and

$$w_{\text{fail}} - w_{\text{succ}} = \lambda N \frac{1 - e^{-1/\tau}}{F + Se^{-1/\tau}} > 0. \tag{9}$$

Proof. For a failed relation, $\exp(-s_i/\tau) = 1$, whereas for a successful relation it equals $e^{-1/\tau} < 1$. Since $F + Se^{-1/\tau} < F + S = N$ when $S > 0$, $w_{\text{fail}} > 1$. Similarly, $Ne^{-1/\tau} < F + Se^{-1/\tau}$ when $F > 0$, which gives $w_{\text{succ}} < 1$. Their difference follows directly by subtraction. □

Bottleneck-weighted advantage. The original BPO advantage is reweighted as

$$\widetilde{A}_i = w_i A_i^{\text{BPO}}. \tag{10}$$

The resulting objective is

$$\mathcal{L}_{\text{BBPO}} = -\frac{1}{N} \sum_{i=1}^{N} \min\left[\rho_i(\theta)\widetilde{A}_i, \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\,\widetilde{A}_i\right] + \beta \widehat{\mathcal{L}}_{\text{KL}}. \tag{11}$$

where

$$\widehat{\mathcal{L}}_{\text{KL}} = \frac{1}{N+1} \sum_{i=0}^{N} \left[\log \pi_\theta(y_i \mid v_i, q) - \text{sg}(\log \pi_{\text{ref}}(y_i \mid v_i, q))\right]. \tag{12}$$

π_{ref} denotes the reference branch, and $\text{sg}(\cdot)$ stops gradients through its score. This sampled log-ratio regularizer limits policy drift while allowing the shared multimodal representation to adapt during post training. At each update, the reference branch shares the current visual and object-state representation with the policy but evaluates the language decoder with the task-specific LoRA adapter disabled. Its score is detached from the computational graph. The reference therefore evolves with the learned visual representation across training steps while providing a fixed comparison target within each update. Hence, BBPO is a credit-allocation mechanism rather than an auxiliary consistency loss or additional supervision objective. Conceptually, BPO specifies how behavior should respond across evidence interventions, whereas BBPO additionally reallocates optimization effort according to which relations within the current pack remain unsatisfied.

3.4 Training Pipeline

Our training pipeline contains three stages. **SFT-A** establishes single-visit perception using object inventory, category, binary state, and temporal-visibility supervision. **SFT-B** introduces paired-revisit reasoning and is balanced across five logical targets: proved false-to-true, proved true-to-false, proved no-change, missing Visit-1 state, and missing Visit-2 state. The main checkpoints are

trained for 10 SFT-B epochs with 25% SFT-A replay. **Post-training** then initializes either BPO or BBPO from the same seed-specific SFT-B checkpoint and trains for two further epochs.

During post-training, each case is expanded into a Visit Pack or Revisit Pack according to Sec. 3.1. The policy produces one rollout for each selected view, and the verifier evaluates both view-level correctness and the required behavioral contract relative to the identity anchor. Standard BPO and BBPO use the same data and training settings; they differ only in the allocation of non-anchor advantages: BPO: A_i^{BPO}, BBPO: $w_i A_i^{\mathrm{BPO}}$.

4 EXPERIMENTS

Data. We derive our revisit-reasoning data from the simulated portion of M^3-Verse, in which each scene contains two temporally ordered observations of the same environment: Visit 1 is recorded before an environment modification and Visit 2 after it. The modification itself is not observed. We perform a one-time observability audit over all 270 scenes using RGB video, simulator state annotations, agent trajectories, and instance segmentation. We retain object identities shared across visits and consider seven binary state properties: open/closed, whole/sliced, intact/broken, off/on, clean/dirty, uncooked/cooked, and empty/filled-with-liquid. We process **M3R-5** and **M3R-3** as our benchmarks, which require the target to occupy at least 5% and 3% of the image in two frames per visit, under same scene-disjoint split of 216/27/27 scenes for training/validation/test. Detail of data processing is discussed in Appendix C.

Since frame sampling is not the focus in this work, behavioral views are materialized lazily at 16 frames per visit and 1 fps, whereas SFT-A uses 32 frames per visit. Evidence-incomplete examples supervise the answer target without forcing privileged state or transition heads to reconstruct deliberately removed evidence.

Model and trainable parameters. Our main model uses a Qwen3.5-2B multimodal backbone augmented with the query-conditioned target-state module described in Sec. B. Rank-16 LoRA with $\alpha = 16$ and dropout 0.05 is applied to language-model linear layers. The final two native vision blocks and the complete target-state/transition module are trainable, while the remaining backbone and native visual merger are frozen. This gives 55.87M trainable parameters for the 2B model. During behavior-pack post-training, we use learning rates of 3×10^{-6} for language LoRA, 1×10^{-6} for the final vision blocks, and 5×10^{-5} for the target-state module. Rollouts use temperature 0.7 and top-p 0.9, while evaluation is deterministic. All results are averaged across 3 different seeds.

Reference-policy variants. We evaluate both frozen and adaptive reference designs. The *frozen reference* holds the reference policy fixed throughout behavior-pack post-training. The *adaptive reference* shares the current visual and target-state representation with the policy while evaluating the language decoder with the task-specific LoRA adapter disabled; its reference score is detached within each update. Our implementation uses an on-policy specialization of the general objective in Sec. 3.3. The rollout-policy log-probability in the nominal importance ratio is the stop-gradient copy of the current-policy log-probability, thus the forward importance ratio is one and PPO clipping is inactive. Standard BPO and BBPO use the same update and reference implementation within each controlled comparison; BBPO differs only through the bottleneck-aware advantage scaling.

Evaluation metrics. Our primary metric is *all-view accuracy*: a Visit or Revisit Pack is counted as correct only when all required view-level answers and behavioral relations are satisfied. We additionally report identity accuracy as a control for ordinary evidence-complete performance and three contract-aligned diagnostics: proved-transition accuracy for sensitivity, nuisance invariance for stability, and missing-premise diagnosis for abstention.

4.1 MAIN RESULTS

Table 1 compares the SFT-B checkpoint before posting training with BPO and BBPO under frozen and adaptive references. Despite relatively strong performance on individual diagnostics, including 83.9% proved-transition accuracy, the SFT-B model achieves only 14.2% overall all-view accuracy, with particularly low Visit-Pack all-view accuracy (4.6%). This large gap indicates that supervised

Table 1: Main results on M3R-5. AV denotes all-view accuracy.

Ver.	Method	Overall AV	Revisit AV	Visit AV	Revisit ID	Visit ID	Transition	Nuisance	Missing
0	SFT-B zero-shot	14.2±0.5	33.3±1.6	4.6±0.8	80.5±1.6	79.9±0.8	83.9±0.8	83.3±4.3	79.9±1.8
I	mask intervention + BBPO	41.4±2.5	12.6±3.3	55.7±3.5	71.3±1.6	82.2±0.8	75.9±2.4	71.3±3.3	55.7±5.9
II	BPO + frozen ref.	62.8±1.4	52.9±4.3	67.8±0.8	87.4±4.3	**82.8±0.0**	82.2±2.9	85.6±4.3	91.1±3.3
III	BPO + adaptive ref.	63.6±1.1	54.0±4.3	68.4±0.8	**89.7±2.8**	82.2±0.8	85.1±3.5	86.8±2.2	91.4±2.5
IV	BBPO + frozen ref.	64.8±1.1	55.2±2.8	69.5±0.8	**89.7±2.8**	82.2±0.8	84.5±2.4	86.8±0.8	91.1±2.0
V (Full)	**BBPO + adaptive ref.**	**66.3±0.5**	**57.5±3.3**	**70.7±1.4**	**89.7±2.8**	82.2±0.8	**85.6±1.6**	**88.5±0.8**	**91.4±1.9**

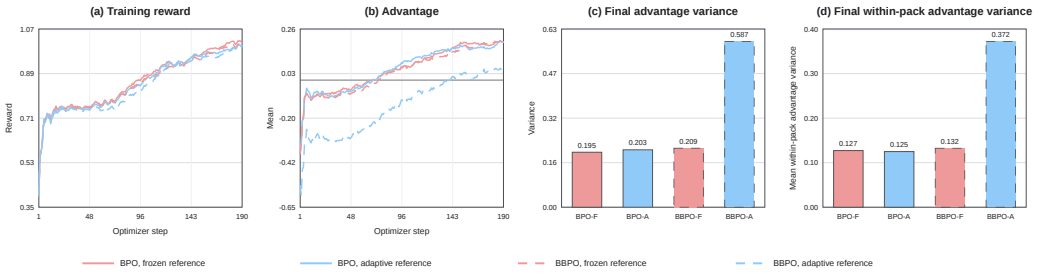

Figure 3: Training dynamics under BPO and BBPO with frozen (F) and adaptive (A) references. (a) Training reward and (b) mean anchor-relative advantage over optimizer steps. (c) Final variance of anchor-relative advantages across training examples. (d) Final mean within-pack advantage variance, computed by first measuring the variance of non-anchor advantages within each behavior pack and then averaging across packs.

reasoning over individual observations does not by itself yield consistent behavior across coordinated interventions. Post training substantially closes this gap: adaptive-reference BPO reaches 63.6% overall all-view accuracy, while our full adaptive-reference BBPO further improves it to 66.3%, with gains on both Revisit Packs ($54.0\% \rightarrow 57.5\%$) and Visit Packs ($68.4\% \rightarrow 70.7\%$). The identity accuracies remain unchanged between adaptive BPO and BBPO, suggesting that the improvement comes from stronger cross-view behavioral consistency rather than sacrificing ordinary evidence-complete performance. We additionally compare the mask-based intervention of original BPO-style as Ver I, and more discussions are provided in Appendix Sec. D.2. Overall, adaptive-reference BBPO gives the superior primary all-view behavioral results.

Training dynamics and within-pack differentiation. Fig. 3 compares the optimization dynamics of BPO and BBPO under frozen and adaptive references. All four variants achieve similar training reward, whereas BBPO with an adaptive reference maintains a noticeably lower mean anchor-relative advantage over most of training. Importantly, this behavior is accompanied by substantially greater advantage dispersion: its final global advantage variance reaches 0.587, compared with 0.195–0.209 for the other variants, while its mean within-pack advantage variance reaches 0.372, versus only 0.125–0.132. This effect is not caused by reference adaptivity alone, since BPO with an adaptive reference exhibits variance comparable to its frozen-reference counterpart, nor by BBPO alone, since frozen-reference BBPO remains similarly low. The increased dispersion is observed only for the joint BBPO + adaptive-reference configuration. Together with the higher all-view accuracy of BBPO with adaptive reference, these results are consistent with stronger differentiation among intervention-dependent behaviors within each pack rather than a uniform increase in per-view advantage. Note that higher variance is not itself desirable; its relevance here lies in its association with improved pack-level behavioral correctness. This behavior is also consistent with Eq.10: because failed relations receive larger w_i, BBPO does not uniformly maximize the mean anchor-relative advantage, but reallocates update magnitude across relations.

4.2 Scaling and Training Maturity

Scaling to a 4B backbone. We evaluate whether the behavioral advantage of BBPO persists when scaling the multimodal backbone from Qwen3.5-2B to Qwen3.5-4B. Except that $\tau = 1, \lambda = 0.35$, 4B configuration follows the same settings as 2B backbone, with 72.70M trainable parameters.

Table 2: Cross-backbone replication on M3R-5.

Backbone	Method	Overall AV	Revisit AV	Visit AV	Overall Δ
Qwen3.5-2B	BPO	63.60±1.1	54.00±4.3	68.40±0.8	–
Qwen3.5-2B	BBPO	66.30±0.5	57.50±3.3	70.70±1.4	+2.7
Qwen3.5-4B	BPO	64.75±3.6	58.62±7.5	67.82±2.2	–
Qwen3.5-4B	BBPO	**67.43±3.0**	**62.07±7.5**	**70.11±0.8**	+2.7

Table 3: Effect of SFT-B checkpoint maturity on subsequent behavior-pack post training.

SFT-B epoch	BPO AV	BBPO AV	Δ AV	Δ Revisit	Δ Visit
6	61.3	62.1	+0.8	+6.9	-2.3
8	62.1	64.8	+2.7	+3.4	+2.3
10	63.6	**66.3**	+2.7	+3.4	+2.3

Table 2 shows that BBPO reproduces the same mean all-view improvement on the 4B backbone: +2.7 pp overall, +3.4 pp on Revisit Packs, and +2.3 pp on Visit Packs. The absolute BPO and BBPO accuracies both increase by approximately 1.1 pp relative to the 2B setting. BBPO on 4B backbone improves overall all-view accuracy, providing cross-backbone evidence that the behavioral effect is not specific to the 2B policy. We also observe that, with a stronger backbone, relation failures become sparser within a pack. At a fixed λ, BBPO therefore automatically assigns greater amplification to each remaining failure. Proposition A.2 provides a possible explanation for why stronger policies may tolerate smaller λ, but establishing this relationship requires controlling τ and directly measuring within-pack failure density.

Table 4: Results on the more difficult M3R-3 observability regime. BBPO uses $\lambda = 0.35$.

Method	Overall AV	Revisit AV	Visit AV	Revisit ID	Visit ID	Pairwise Rank	Exact Rank	Set F1
BPO	54.92	47.62	58.57	81.90	**74.29**	77.92	64.76	69.60
BBPO	**56.83**	47.62	**61.43**	81.90	71.43	**78.84**	**66.55**	**70.52**

Effect of SFT-B maturity. In Table 3, we also vary the SFT-B initialization to determine whether BBPO remains useful as the supervised revisit reasoner becomes stronger. From epoch 8 to epoch 10, both BPO and BBPO improve by 1.5 pp in absolute all-view accuracy, while the BBPO treatment effect remains unchanged at +2.7 pp. This indicates that BBPO continues to provide additional benefit as the SFT-B initialization becomes stronger rather than merely compensating for an under-trained supervised model. The epoch-10 checkpoint also yields a positive overall BBPO improvement in every matched run and is therefore used for the main experiments.

Robustness to smaller visual targets. We evaluate the same reasoning protocol under M3R-3, which lowers the target-occupancy threshold from 5% to 3%. As shown in Table. 4, BBPO remains beneficial under weaker target observability, improving overall all-view accuracy by 1.9 pp and Visit-Pack all-view accuracy by 2.9 pp. Generally, BBPO remains active under more difficult visual evidence.

5 Conclusion

We study ontic–epistemic reasoning across video revisits, where a model must distinguish what physically changed from what can be justified under perturbed evidence. The proposed BBPO reallocates BPO credit toward unsatisfied within-pack behavioral relations. Across 2 backbones, BBPO consistently improves all-view behavioral correctness over standard BPO, with the strongest results obtained using an adaptive reference and a moderate bottleneck coefficient.

Limitation. Our work uses simulated indoor scenes, seven binary object-state properties, and multiple-choice outputs; generalization to real-world videos, continuous state changes, and repeated same-type objects across visits are not studied.

AI USE STATEMENT

In this work, we used generative AI tools for brainstorming, code implementation, data analysis, report preparation, result double check, and writing polishing. We have not used generative AI tools for idea generation and writing organization. We have reviewed all AI-assisted work.

ETHICS STATEMENT

No Ethics statement is needed.

REPRODUCIBILITY STATEMENT

We explicitly described method and experiment details in Sec. 3, Sec. 4, Appendix, and Supplementary Material, which are sufficient to reproduce the results.

REFERENCES

Luca Annese, Sabrina Patania, Silvia Serino, Tom Foulsham, Silvia Rossi, Azzurra Ruggeri, and Dimitri Ognibene. Who sees what? structured thought-action sequences for epistemic reasoning in llms. In *International Conference on Social Robotics*, pp. 387–399. Springer, 2025.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026.

Adi Gabay, Gabriel Stanovsky, and Liat Peterfreund. Beyond memorization: Distinguishing between pattern-based and epistemic reasoning in llms using epistemic puzzles. *arXiv preprint arXiv:2603.21350*, 2026.

Davide Grossi, Andreas Herzig, Wiebe Van Der Hoek, and Christos Moyzes. Non-determinism and the dynamics of knowledge. In *26th International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pp. 1053–1059, 2017.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514. IEEE, 2024.

Zhaolu Kang, Shiyu Liu, Tailong Luo, Wei Zhang, Yingjie He, Lei Wei, Guansu Wang, Liang He, Siheng Wang, Guangyuan Dong, Jiaqi Su, Shuang Chen, Haoyu Ji, Qishi Zhan, and Kaiyue Zhou. Behavior pack optimization for video MLLM post-training. 2026.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel Bell. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=OkHC30LLpO.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Nishanth Madhusudhan, Vikas Yadav, and Alexandre Lacoste. Knowing when not to answer: Evaluating abstention in multimodal reasoning systems. *arXiv preprint arXiv:2604.14799*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Fnu pramono, John Cai, and Sourabh Kulkarni. TRAPSBench: Vision-language models encode but fail to express epistemic restraint. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=o5cFzLUOf0.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. Egomemreason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. *arXiv preprint arXiv:2605.09874*, 2026.

Kewei Wei, Bocheng Hu, Jie Cao, Xiaohan Chen, Zhengxi Lu, Wubing Xia, Weili Xu, Jiaao Wu, Junchen He, Mingyu Jia, et al. $m\hat{3} - verse$: A" spot the difference" challenge for large multi-modal models. *arXiv preprint arXiv:2512.18735*, 2025.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European conference on computer vision*, pp. 148–166. Springer, 2022.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Sihyun Yu, Nanye Ma, Pinzhi Huang, Hyunseok Lee, Shusheng Yang, June Suk Choi, Ellis Brown, Oscar Michel, Boyang Zheng, Jinwoo Shin, et al. Benchmarking visual state tracking in multimodal video understanding. *arXiv preprint arXiv:2606.03920*, 2026.

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32903–32914, 2026.

Xinyu Zhou, Chang Jin, Carsten Eickhoff, Zhijiang Guo, and Seyed Ali Bahrainian. When silence is golden: Can LLMs learn to abstain in temporal QA and beyond? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PhUCxfS0yf.

Yanxu Zhu, Shitong Duan, Xiangxu Zhang, Jitao Sang, Peng Zhang, Tun Lu, Xiao Zhou, Jing Yao, Xiaoyuan Yi, and Xing Xie. Mohobench: Assessing honesty of multimodal large language models via unanswerable visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29205–29213, 2026.

A PROPOSITIONS

Proposition A.1 (Mean-preserving credit reallocation)**.** *Let $E(\mathcal{B})$ contain N non-anchor relations and let*

$$w_i = (1-\lambda) + \lambda N \frac{\exp(-s_i/\tau)}{\sum_{j\in E(\mathcal{B})} \exp(-s_j/\tau)}, \qquad i \in E(\mathcal{B}), \tag{13}$$

with $\lambda \in [0,1]$ and $\tau > 0$. Then $\sum_{i\in E(\mathcal{B})} w_i = N$, and hence $\frac{1}{N}\sum_i w_i = 1$. Moreover, $1-\lambda \le w_i \le 1-\lambda+\lambda N$, and $\lambda = 0$ recovers the unweighted BPO update.

Proof. Summing the weights over $E(\mathcal{B})$ gives

$$\sum_i w_i = N(1-\lambda) + \lambda N \frac{\sum_i \exp(-s_i/\tau)}{\sum_j \exp(-s_j/\tau)} = N. \tag{14}$$

The bounds follow because the normalized exponential term lies in $[0,1]$, and setting $\lambda = 0$ yields $w_i = 1$ for every relation. □

Table 5: Statistics of the two observability regimes. M3R-3 differs from M3R-5 only in the minimum target-occupancy threshold.

Dataset	Threshold	Scenes	SFT-A	SFT-B	Visit Pack	Revisit Pack
M3R-5	5%	270	4,472	500	310	154
M3R-3	3%	270	5,000	595	362	180

Proposition A.2 (Failure-density-dependent bottleneck amplification)**.** *Assume binary relation-success scores $s_i \in \{0,1\}$ and let $r = F/N$ denote the fraction of failed relations in a pack. Define $a = \exp(-1/\tau)$. Under BBPO, every failed relation receives weight*

$$w_{\text{fail}}(r) = 1 + \lambda \left[\frac{1}{a + (1-a)r} - 1 \right]. \tag{15}$$

For any $\lambda > 0$ and $\tau > 0$,

$$\frac{\partial w_{\text{fail}}}{\partial r} = -\lambda \frac{1-a}{[a + (1-a)r]^2} < 0. \tag{16}$$

Thus, as failures become sparser within a pack, BBPO automatically assigns greater weight to each remaining failed relation.

B Architectural Detail

Compared with the native Qwen3.5 multimodal backbone, we add a lightweight query-conditioned target-state pathway tailored to revisit reasoning, as illustrated in Fig. 2. We tap the pre-merger visual tokens from each visit and condition a small set of learned object queries on the target object, queried property, and question; spatial cross-attention then pools target-relevant evidence into visit-specific representations. These representations are supervised with auxiliary binary state heads and a three-way inter-visit transition head, and a bounded residual writer projects the resulting state and relation features back into the native Qwen visual-token stream before language decoding. The underlying Qwen vision encoder, visual merger, and language decoder are otherwise preserved, with only the final two vision blocks, the added target-state module, and LoRA adapters trained; explicit cross-visit Sinkhorn matching, mask-grounding supervision, and temporal-tracklet supervision are not used in our configuration.

C Detail of Experimental Settings

Data Processing. During data processing, a candidate is retained only when, in each visit, it is trajectory-visible in at least two frames, its segmentation occurs in at least two frames, and at least two frames satisfy the observability regimes, which are set to be at least 5% for **M3R-5** and 3% for **M3R-3**. The distribution of data for each training stage is presented in Table. 5. We split the 270 scenes into 216/27/27 train/validation/testing scenes, which yields 54 validation packs and 33 testing packs. Since the total number is still limited, we combine validation and testing sets (summed to 87 packs for **M3R-5**) for quantitative reports. We intentionally balance transition classes, state configurations, object/property categories, and answer-option positions during data construction to reduce label-frequency and answer-position shortcuts. Hyperparameters and model variants, including λ, reference design, and checkpoint selection, are selected using validation data only; the test split is used once for final reporting.

Hardware and Seeds. All experiments use 4× NVIDIA H20 (most results) or 4× NVIDIA Pro 6000 (results using 4B backbone as in Table 2). All runs use seeds $\{42, 43, 44\}$.

Visit and Revisit Packs. The identity view retains the complete visual evidence and serves as the anchor. Reversing visits swaps the two complete observations without reversing frame order within either visit, so directional transitions reverse while no-change remains unchanged. The nuisance controls modify temporal content while retaining certified target evidence and should preserve the identity answer. Target-frame withholding removes every frame containing target pixels from one visit, whereas target-free subclip retains only the longest contiguous interval in which the target

Table 6: Effect of the bottleneck coefficient λ on M3R-5.

λ	Overall AV	Revisit AV	Visit AV
0.00	64.37	52.87	70.11
0.25	64.37	52.87	70.11
0.50	64.37	54.02	69.54
0.70	**66.28**	**57.47**	70.69
1.00	64.37	50.57	**71.26**

Table 7: Effect of replacing spatial masking with natural temporal evidence withholding. Results are mean ± seed standard deviation (%). The largest improvements occur on evidence-removal views, where a mask-specific shortcut would be expected to help most.

Metric	Mask-based	Temporal withholding	Δ
Revisit identity	71.3±1.6	89.7±2.8	+18.4
Visit identity	82.2±0.8	82.2±0.8	+0.0
Overall all-view accuracy	41.4±2.5	66.3±0.5	+24.9
Proved transition	75.9±2.4	85.6±1.6	+9.7
Nuisance invariance	71.3±3.3	88.5±0.8	+17.2
Missing-premise diagnosis	55.7±5.9	91.4±1.9	+35.7
Revisit-Pack all-view accuracy	12.6±3.3	57.5±3.3	+44.9
Revisit target removal, Visit 1	51.7	94.3	+42.6
Revisit target removal, Visit 2	55.2	85.1	+29.9
Revisit target-free view, Visit 1	60.9	96.6	+35.7
Revisit target-free view, Visit 2	55.2	89.7	+34.5
Visit-Pack target removal	81.6	98.9	+17.3
Visit-Pack target-free view	83.3	98.3	+15.0

never appears. These evidence-incomplete views require the model to identify the missing visit-state premise and reason over the transitions compatible with the remaining evidence. Every transformed visit must retain at least 16 source frames.

D More Experiments

D.1 Effect of Bottleneck Strength

The coefficient λ controls how strongly BBPO reallocates optimization pressure from uniform BPO credit toward currently unsatisfied relations. We ablate $\lambda \in \{0, 0.25, 0.5, 0.7, 1.0\}$ on M3R-5 while keeping the training protocol fixed. Table 6 shows that the setting $\lambda = 0.7$ achieves the highest overall all-view accuracy at 66.28% and is best or tied for best in every run.

The improvement is driven primarily by Revisit-Pack reasoning. Revisit all-view accuracy increases from 52.87% at $\lambda = 0$ to 57.47% at $\lambda = 0.7$. The result also supports the mixture interpretation of BBPO: retaining a nonzero uniform component is important, while excessive bottleneck concentration can over-focus a subset of relations and weaken the all-view Revisit-Pack contract.

D.2 Effect of Natural Evidence Withholding

To test whether performance depends on the spatial masks used in the original BPO-style intervention construction, we compare a **mask-based** setting (also referred to as Ver I in Table 1) with our **temporal-withholding** setting in Table 7. The former removes target evidence using target-aligned spatial masks, whereas the latter contains no artificial masks: target evidence is removed either by deleting every target-containing frame or by retaining a naturally occurring target-free temporal subclip. All results are averaged over three matched seeds.

The temporal-withholding construction substantially improves behavioral consistency despite removing the most obvious intervention cue: the target-aligned spatial mask. Overall all-view accuracy increases by 24.9 pp and Revisit-Pack all-view accuracy by 44.8 pp, while missing-premise diagnosis improves from 55.7% to 91.4%. More importantly, the largest gains occur precisely on the evidence-removal views for which a mask-based shortcut would be most plausible. For example,

withholding Visit 1 improves from 51.7% under masking to 94.3% under frame withholding, while the corresponding natural target-free view reaches 96.6%. Since the temporal construction contains no artificial mask pixels and answer-option positions are shuffled, these results provide evidence against a fixed spatial-mask or answer-position shortcut.

However, the experiment does not rule out all forms of intervention recognition. A model may still learn that a target-absent or shortened clip indicates a missing premise without fully recovering the underlying object state. Our matched non-target-frame controls reduce the likelihood of a trivial duration or frame-count heuristic, because they remove temporal content while preserving the target evidence and require the original answer to remain unchanged.