TARA: Task-Aware Reasoning with Temporal Evidence Amplification for BlackSwan MCQ

Xingyu Zhu[1,2], Yu Zhang[1], Wenwu He[1], Yuyang Sun[3], Haoxuan Ma, Yongliang Wu[3], Xiaogang Wang[4], Yuxia Chen, Zhenxiang Jiang, Yangguang Ji, Wenbo Zhu, Xu Yang[3]
[1]Beijing Freedo Technology Co., Ltd. [2]National University of Singapore
[3]Southeast University [4]Southwest University

Abstract

The BlackSwan Challenge evaluates whether video-language models can reason about unexpected events rather than merely describe common visual patterns. We present TARA, a training-free inference framework for BlackSwan multiple-choice reasoning. TARA reformulates each question as hypothesis-conditioned video reasoning and combines five modules: task-aware evidence routing, temporal evidence amplification, causal state transition reasoning, option-level hypothesis arbitration, and adaptive inference control. The framework explicitly separates the two benchmark regimes: Detective questions require inferring a hidden middle event from pre/post clips, while Reporter questions require verifying the surprising event from the full three-clip video. On the validation split, our native-speed Gemini 3.1 Pro Preview baseline achieves 75.2% answered accuracy. A diagnostic slow-motion retry on validation failures shows that a non-trivial subset of errors is perception-sensitive, motivating our final label-free temporal amplification rule for short clips.

1. Introduction

Recent video-language models are increasingly strong at describing visible actions, objects, and scenes, but this does not necessarily imply robust event-level reasoning. In real videos, the most important moment is often brief, unexpected, and causally decisive: a slip, a collision, a failed landing, a sudden grab, or a small object interaction. Understanding such events requires more than captioning what is visible; it requires identifying which observation changes the state of the scene and why.

The BlackSwanSuite benchmark [1] is designed around this gap. Its multiple-choice setting is especially challenging because the answer options are often near-miss hypotheses rather than unrelated labels. Several choices may mention the same actors and objects, while differing only in the direction of contact, the success or failure of an action, the hidden cause of a later state, or the exact outcome of the surprising event. As a result, a coarse video summary can be insufficient even when it is visually plausible.

This challenge becomes more pronounced in the two official task regimes. In *Detective*, the model sees only the pre-event and post-event clips and must infer the missing middle event from a state transition. In *Reporter*, the model sees the full video and must use the event clip to select the option most consistent with the actual surprising moment. These two regimes require different uses of evidence, yet a naive zero-shot pipeline tends to treat both as ordinary video QA.

We therefore propose **TARA**, short for Task-Aware Reasoning with Temporal Evidence Amplification. TARA is an inference-time framework that makes the benchmark structure explicit: it routes visual evidence according to the task, exposes short transient events through temporal amplification, frames Detective examples as causal state-transition problems, and treats the MCQ options as competing hypotheses to arbitrate against the video evidence. This keeps the system lightweight while aligning the reasoning process more closely with the structure of BlackSwan MCQ.

2. Method

This section details the TARA pipeline. Given a BlackSwan MCQ sample, TARA proceeds in three stages: it routes visual evidence according to the task type, amplifies short transient events through dense temporal sampling, and arbitrates among the candidate options via hypothesis-conditioned verification. The overall pipeline is illustrated in Figure 1.

2.1. Task-Aware Evidence Routing

BlackSwan defines two reasoning regimes with different information constraints. TARA therefore constructs different video inputs for Detective and Reporter questions.

For Detective, the model receives only the pre-event and post-event clips, labeled as "BEFORE the event" and "AFTER the event". The model must infer the hidden middle event that best explains the observed transition. This is an abductive reasoning setting.

For Reporter, the model receives the full sequence: pre-event, event, and post-event. The middle clip is explicitly

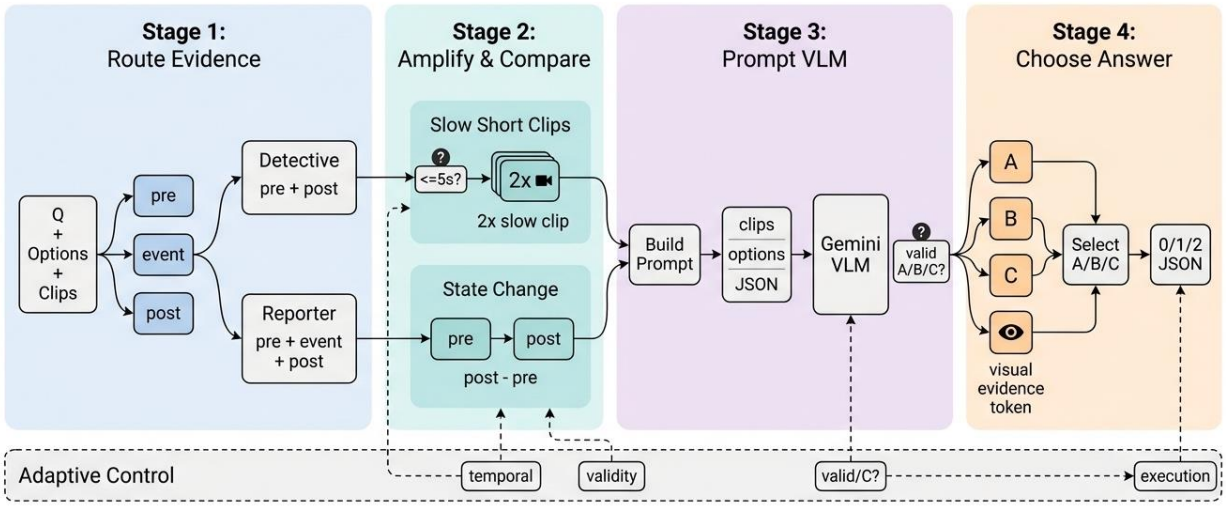

Figure 1. Overview of TARA. The pipeline converts BlackSwan MCQ from generic video QA into task-aware hypothesis verification.

labeled as "THE EVENT", allowing the model to verify which answer choice best describes the surprising moment. This is a defeasible reasoning setting.

This module prevents the two tasks from collapsing into a generic video-QA format. The visual evidence remains aligned with the official benchmark constraints.

2.2. Temporal Evidence Amplification

Many BlackSwan questions depend on short-lived evidence, such as a collision, fall, bite, grab, slip, failed landing, ball impact, or object breakage. These events may occupy only a few frames. If the video model misses these frames, it may understand the scene but still choose the wrong causal explanation.

We introduce a temporal evidence amplification module that applies frame-preserving 2x slowdown to short clips. The operation does not synthesize new content or alter spatial evidence; it stretches the original frames over a longer temporal window so that transient causal evidence receives more attention. At test time, we use a label-free adaptive rule: clips with duration at or below 5 seconds are amplified, while longer clips remain unchanged.

2.3. Causal State Transition Reasoning

Detective questions are not simply about describing two clips independently. The core problem is to explain why the post-event state follows from the pre-event state. TARA therefore formulates Detective reasoning as state transition explanation: the model compares people, objects, poses, spatial relations, and result states before and after the missing event, then selects the hypothesis that best explains the change.

This reformulation moves the model from "what is visible in each clip" to "what hidden cause explains the transition". For example, a person becoming prone, an object becoming broken, or two entities becoming separated should be interpreted as the result of an intervening event rather than as isolated observations.

2.4. Option-Level Hypothesis Arbitration

The three MCQ options are treated as competing event hypotheses. TARA performs option-level hypothesis arbitration by asking the model to compare each option against the same visual evidence and choose the most consistent explanation.

The prompt contains ordered clip labels, the official question, the three answer options, and a constrained JSON schema:

```
{"answer": "A",
"reason": "brief explanation"}
```

The answer field is used for scoring, while the rationale helps validation-time error analysis. This design is more suitable than open-ended description because it forces evidence matching within the provided hypothesis set.

2.5. Adaptive Inference Control

TARA uses an adaptive inference controller to manage sample-specific processing. The controller includes three policies. First, the temporal policy decides whether to apply 2x slowdown based on clip duration. Second, the validity policy checks whether the model output is a valid A/B/C response and flags malformed answers for retry or

forced-choice post-processing. Third, the execution policy reuses cached videos and retries transient API failures. These mechanisms make the inference path adaptive without changing the underlying model.

3. Experiments and Analysis

3.1. Evaluation Protocol

We evaluate TARA under the official BlackSwan multiple-choice setting. Each sample contains a question and three answer options, and the system must output one of A/B/C. The final prediction is converted to the official zero-indexed label format for submission. We report overall accuracy as well as the two official task-specific metrics, Detective accuracy and Reporter accuracy.

Detective and Reporter examples are evaluated with different visual evidence constraints. For Detective, the model receives only the pre-event and post-event clips and must infer the missing event. For Reporter, the model receives the pre-event, event, and post-event clips and must verify the visible surprising event. This separation is kept throughout validation and hidden-test inference.

3.2. Implementation Details

Our final system uses Gemini 3.1 Pro Preview [2] as the base video-language model. We use deterministic generation with temperature 0 and a structured multiple-choice prompt. For each example, TARA constructs task-specific clip inputs, applies temporal evidence amplification when the clip is short, and asks the model to return a constrained JSON object containing the selected option and a brief rationale. The option field is used for scoring, while the rationale is used only for error analysis.

At test time, temporal amplification is triggered by a label-free rule: clips with duration at or below 5 seconds are slowed down by 2x, while longer clips are kept at native speed. This rule is selected from validation diagnostics and does not use hidden-test labels.

3.3. Validation Results

Table 1 reports validation performance for the native-speed inference baseline before applying the final adaptive temporal rule. The system achieves 75.2% answered accuracy overall. Reporter accuracy is higher than Detective accuracy, showing that verifying a visible event is easier than reconstructing a hidden event from before/after evidence.

3.4. Temporal Amplification Diagnostic

To understand whether errors are caused by reasoning failure or missed transient visual evidence, we performed a diagnostic 2x slowdown retry on validation examples that were initially wrong or failed to parse. This analysis is used only to identify error modes. It is not an oracle selection strategy, since validation labels are used to choose the failed subset.

Group	Correct / Answered	Accuracy (%)
Overall	1346 / 1789	75.2
Detective	827 / 1159	71.4
Reporter	519 / 630	82.4
Detective-easy	385 / 521	73.9
Detective-medium	368 / 513	71.7
Detective-hard	74 / 125	59.2
Reporter-easy	170 / 194	87.6
Reporter-medium	197 / 247	79.8
Reporter-hard	152 / 189	80.4

Table 1. Validation results with native-speed inference. Accuracy is computed on parsed answers.

Diagnostic quantity	Value
Retried wrong/failed validation cases	440
Recovered correct cases after 2x slowdown	123
Recovery rate	28.0%

Table 2. Validation-only temporal amplification diagnostic. The diagnostic motivates the test-time duration rule but is not used as an oracle during hidden-test evaluation.

Metric	Accuracy (%)
Overall Accuracy	76.28
Detective Accuracy	72.57
Reporter Accuracy	82.26

Table 3. Official hidden-test results of the final TARA submission.

Among 440 retried validation cases, 123 became correct after slowdown, corresponding to a recovery rate of 28.0%. This suggests that a meaningful portion of BlackSwan errors are perception-sensitive: the model may select the correct causal hypothesis once the short decisive event is easier to inspect. This finding motivates the final label-free temporal amplification rule used for hidden-test inference.

3.5. Official Hidden Test Results

Table 3 reports the official hidden-test performance of our final TARA submission. The system achieves 76.28% overall accuracy, with 72.57% on Detective and 82.26% on Reporter. The same trend as validation is observed: Reporter remains easier because the decisive event is directly visible, while Detective requires abductive inference from state changes.

3.6. Analysis

The validation and hidden-test results show a consistent gap between Detective and Reporter. This supports the design

choice of separating the two task regimes instead of treating BlackSwan as generic video QA. Reporter questions benefit from direct event observation, whereas Detective questions require the model to infer a missing cause from the transition between the pre-event and post-event clips.

The temporal diagnostic further indicates that some failures are not purely semantic. Short events such as collisions, falls, catches, slips, or object impacts can be missed by a single-pass video model, leading to plausible but incorrect options. TARA addresses this failure mode with a simple duration-based amplification policy, keeping the method training-free while improving the visibility of transient evidence.

4. Conclusion

We introduced TARA, a training-free inference framework for BlackSwan MCQ reasoning. By combining task-aware evidence selection, temporal evidence amplification, causal state transition modeling, and option-level hypothesis arbitration, TARA converts video MCQ into structured unexpected-event reasoning. Validation analysis shows that hidden-event inference is the main challenge and that missed short-duration evidence is an important, partially addressable source of error.

References

[1] Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24201–24210, 2025. 1

[2] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3