VEPAGENT: BRIDGING CAUSAL-TRANSITION VIA TOOL-AUGMENTED REINFORCEMENT LEARNING FOR VIDEO EVENT PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have demonstrated remarkable potential in video understanding, yet their reliance on retrospective summarization and text-centric priors often limits their ability to bridge unobserved causal transitions when applied to Video Event Prediction (VEP). To address this, we propose **VepAgent**, an agentic framework that integrates causal-transition reasoning with tool-augmented reinforcement learning (RL) for robust VEP. Unlike prior methods that passively project future trajectories from historical dependencies, our approach explicitly models the logical progression from terminal observed states to future events. Specifically, we first construct *futurebench-4K*, a high-quality chain-of-thought dataset for supervised fine-tuning (SFT) that effectively bridges the causal-logic gap by structuring the deduction of unobserved intermediate states. Subsequently, we develop a diagnostic tool library integrating state tracking, frame retrieval, and region magnification, enabling the agent to dynamically augment reasoning with external tools to recover missing spatio-temporal evidence and resolve visual ambiguities during inference. Moreover, we propose a composite reward mechanism that jointly optimizes prediction accuracy, causal coherence, and reliable prior, compelling the agent to rely on genuine visual grounding rather than superficial textual similarities. Extensive evaluations on FutureBench and NEPBench datasets demonstrate that our method achieves state-of-the-art performance, significantly outperforming larger MLLMs and validating the empirical effectiveness of our agentic, future-oriented reasoning paradigm.

1 INTRODUCTION

Video Event Prediction (VEP) aims to anticipate plausible future events based on partially observed visual evidence (Nako & Jatowt, 2025; Schmidt et al., 2025). This forward-looking capability is foundational for next-generation applications in areas including proactive robotics (Abbate et al., 2024), autonomous driving (Hu et al., 2020), intelligent surveillance (Karim et al., 2022), etc. Unlike standard video understanding tasks that primarily require descriptive summarization of observed content, VEP demands a deeper causal comprehension of temporal dynamics to future trajectories. A central challenge is developing agents that can bridge observed reality with unobserved futures under incomplete visual evidence.

Recent breakthroughs in Multimodal Large Language Models (MLLMs) (Jain et al., 2024; Wu et al., 2024; Li et al., 2024) have advanced both video perception and reasoning. By leveraging vast world knowledge and grounding it in visual evidence, MLLMs (Khattak et al., 2025) interpret nuanced visual details and generate coherent narratives from video inputs with high fidelity. Models like the Qwen-VL series (Bai et al., 2025b;a) have demonstrated strong performance across a spectrum of challenging tasks, including video dialogue, temporal grounding, and few-shot learning.

Despite their capabilities, applying MLLMs to VEP still exposes three fundamental limitations. ❶ **Causal-Logic Gap.** While MLLMs can successfully reconstruct historical event chains, this retrospective summarization does not inherently confer the ability to extrapolate future trajectories. There exists an unobserved intermediate state transition between the terminal frame and the future event that requires structured causal bridging, rather than merely relying on historical dependencies

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.

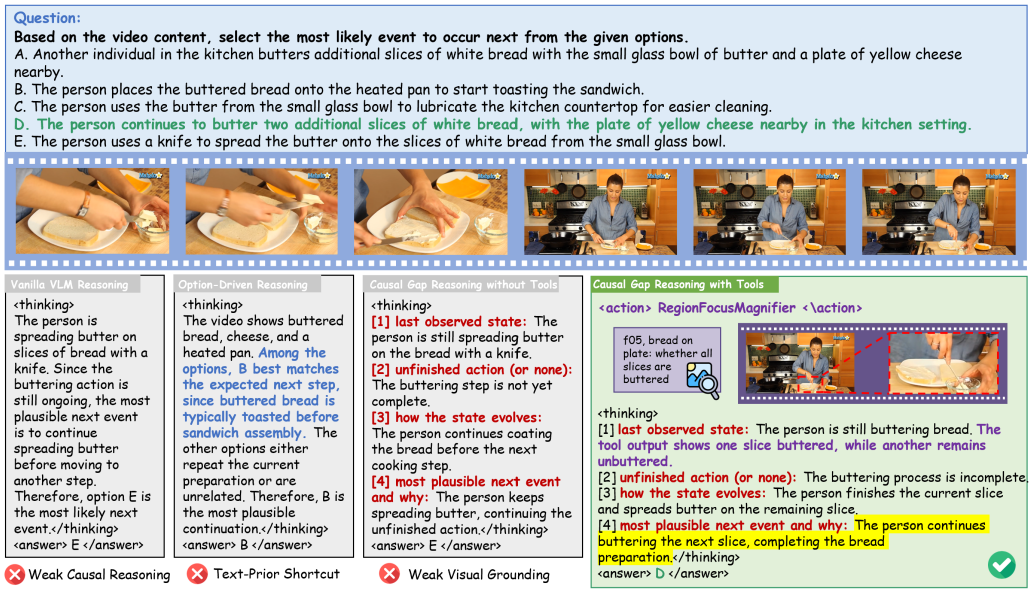

Figure 1: **Key insight of VepAgent.** (a) Vanilla VLM reasoning weakly models the future transition and predicts a coarse continuation of the observed action. (b) Option-driven reasoning is distracted by textual priors and prematurely predicts the cooking stage. (c) Causal-gap reasoning captures the ongoing buttering process, but misses the fine-grained visual state and selects the wrong continuation. (d) With tool-augmented visual evidence, VepAgent identifies the remaining unbuttered slice, correctly reasons over the unfinished action, and predicts the next event.

to project outcomes. ❷ **Language-Prior Shortcut.** Instead of faithfully grounding reasoning in video content, models frequently exploit linguistic priors by selecting the answer option that shares the highest textual similarity with the query. This over-reliance on surface-level lexical patterns and statistical co-occurrences often overrides actual visual states, leading to visually unsupported predictions. ❸ **Visual-Evidence Omission.** Real-world videos frequently suffer from partial occlusions, rapid motion, or suboptimal frame sampling, truncating critical spatio-temporal cues for state verification. Without mechanisms to dynamically recover missing transient actions or fine-grained spatial details, MLLMs are forced to resort to biased estimations in ambiguous scenarios.

To address these challenges, we argue that an effective VEP agent requires three synergistic capabilities. ❶ **Causal-Transition Reasoning**, which explicitly models the logical trajectory from the terminal observed state, through the ongoing action, to the future event, enabling the model to construct physically plausible predictions rather than relying on historical summaries. ❷ **Anti-Prior Reinforcement**, which penalizes reasoning shortcuts that exploit superficial textual similarities between queries and options, thereby encouraging predictions based on actual video states. ❸ **Tool-Augmented Perception**, which dynamically integrates a multimodal tool library to recover omitted visual evidence and magnify fine-grained spatio-temporal details, transforming the static observer into an active perceiver capable of capturing fleeting state transitions. Together, these capabilities reformulate VEP from passive future guessing into an active process of causal analysis, visual verification, and temporal grounding.

Therefore, we introduce **VepAgent**, a framework that unifies causal-transition reasoning with **tool-augmented reinforcement learning (RL)** and an anti-prior reward for video event prediction. The ***futurebench-4K*** chain-of-thought (CoT) dataset for **supervised fine-tuning (SFT)** structures the logical progression from observed states and ongoing actions to future events. It teaches the model to identify unfinished dynamics and physically plausible continuations rather than merely summarizing the past, bridging the causal-logic gap.

To overcome language-prior shortcuts and resolve visual-evidence omission, we further develop a tool-augmented framework tailored to group relative policy optimization (GRPO) in the RL stage. During inference, the agent dynamically interacts with a diagnostic tool library that tracks state tran-

sitions, retrieves frame evidence, magnifies local regions, and extracts fine-grained visual facts. A composite reward jointly optimizes prediction accuracy, causal-gap bridging, and an explicit Anti-Prior penalty, guiding the agent to learn not only *how* to use external tools, but also *when* and *why* additional evidence should be acquired for grounded future predictions. In summary, our contributions are four-fold:

- **Tool-Augmented Agentic RL.** We introduce a tool-augmented agentic RL framework that unifies causal-transition reasoning with dynamic evidence acquisition. Structured SFT and a Causal Gap Reward enable VEP through dynamic, future-oriented reasoning rather than retrospective summarization.
- **Composite Anti-Prior Reward.** We design a composite reward with an Anti-Prior term that penalizes language-shortcut behaviors and encourages reliance on visual evidence rather than superficial textual similarities.
- **Spatiotemporal Multi-Tool Integration.** We build a multimodal tool library integrating state tracking, focused frame retrieval, region magnification, and visual detail extraction, dynamically augmenting reasoning to resolve fine-grained detail ambiguity and key-frame omission.
- **Empirical Effectiveness.** On FutureBench and NEPBench, VepAgent (4B) raises average accuracy by 14.58 and 22.70 percentage points over the prior strongest reported MLLM baselines, from 66.86% to 81.44% and from 47.50% to 70.20%, respectively, under the same evaluation protocols.

2 Related Work

Video Event Prediction. Previous VEP methods primarily originated from script event induction in NLP (Granroth-Wilding & Clark, 2016) and early action anticipation (Girdhar & Grauman, 2021), relying on structured datasets like VidSitu (Sadhu et al., 2021) and EPIC-Kitchens (Damen et al., 2018) to model dynamic event evolution. To capture complex semantic hierarchies, recent works like VidEvent (Liang et al., 2025) and EventFormer (Su et al., 2025) introduced large-scale structured benchmarks and node-graph attention mechanisms for action-centric prediction. Earlier work such as VLEP (Lei et al., 2020) formulated next-event prediction from partially observed videos, while more recent benchmarks like FutureBench (Wang et al., 2026) evaluate future-event reasoning in modern MLLMs. However, native MLLMs (Su et al., 2026; Chen et al., 2025) still suffer from insufficient visual utilization and language-prior shortcuts when extrapolating future trajectories. **VepAgent** combines MLLM backbones with tool-augmented RL to target these failure modes in video event prediction.

Multimodal LLMs Reasoning. Recent advances in Large Language Models (LLMs) have demonstrated that RL-based post-training can enhance reasoning capabilities, as exemplified by Qwen-VL (Bai et al., 2025a), OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), etc. These paradigms have been extended to MLLMs for tasks like mathematical VQA (Peng et al., 2025), image segmentation (Liu et al., 2025), and video understanding (Feng et al., 2026a). However, existing methods struggle with language-prior shortcuts (Chen et al., 2026; Feng et al., 2026b). To address this, we propose causal-transition reasoning via tool-based RL for visual grounding.

Tool-Augmented Agentic System. Early works like FAST (Sun et al., 2025) and MVoT (Li et al., 2025) introduce visual evidence into reasoning, forming multimodal CoT for image tasks. LLaVA-Plus (Liu et al., 2023) pioneered training strategies for tool use, while VPD (Hu et al., 2024) leveraged program-derived data to transfer tool skills. Recent methods like TACO (Ma et al., 2024) and PyVision (Zhao et al., 2025) extend tool use with RL, while VITAL (Zhang et al., 2025) further enables dynamic, on-demand visual evidence acquisition for video reasoning. Building on this line of work, VepAgent contributes VEP-specific causal-transition reasoning, with diagnostic tools designed to recover missing temporal transitions and fine-grained visual evidence.

3 Methodology

Overview. We present VepAgent, a framework that integrates causal-transition reasoning with tool-augmented RL and an anti-prior reward for video event prediction, as shown in Fig. 2. We next detail the problem formulation and workflow, training-data construction, SFT, and GRPO-based RL.

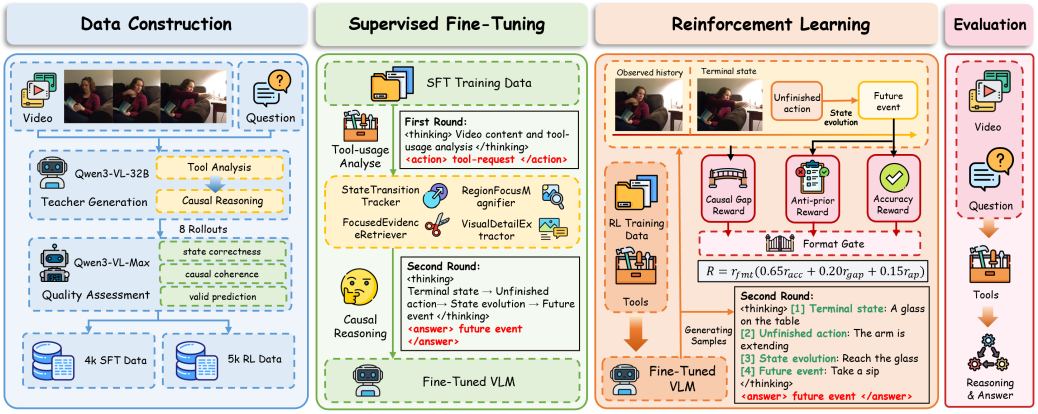

Figure 2: **Pipeline of VepAgent.** (i) Construct tool-augmented causal-transition CoT data by explicitly modeling the progression from terminal observed states to future events. (ii) Supervised fine-tuning initializes tool-use and causal-gap reasoning behaviors. (iii) GRPO optimizes the policy with accuracy, causal-gap, and anti-prior rewards. (iv) The trained agent dynamically integrates diagnostic tools for grounded future-event prediction.

3.1 Problem Formulation & Workflow Design

The goal in VEP is to anticipate the most plausible future event based on partially observed visual evidence. The input is a partially observed video V, a query Q, and a set of candidate future events (options) $\mathcal{O} = \{o_k\}_{k=1}^{K}$, while the output is the predicted future event $y \in \mathcal{O}$. Unlike standard video understanding that relies on retrospective summarization, VEP demands forward-looking causal reasoning. Formally, the model must accurately infer the terminal observed state S_{final} and identify the ongoing, unfinished action A_{ongoing} from V, and then logically bridge the causal-logic gap, which represents the unobserved intermediate state transition, to project the next physically plausible state.

To achieve this goal and resolve the visual-evidence omission during inference, we design the following two-stage agentic workflow to integrate domain-specific diagnostic tools:

Stage 1: Observation & Tool Selection. Given a partially observed video V and query Q, our model first forms an initial estimate of the terminal observed state S_{final} and the ongoing action A_{ongoing}. Concurrently, it constructs a causal hypothesis of the future trajectory. For instances where critical state transitions are ambiguous, occluded, or missed due to uniform frame sampling, our agentic framework formulates a plan by selecting an appropriate subset of tools from a diagnostic tool library $\mathcal{T}$. This library comprises a State Transition Tracker ($\mathcal{T}_{\text{s}}$), a Focused Evidence Retriever ($\mathcal{T}_{\text{f}}$), a Region Focus Magnifier ($\mathcal{T}_{\text{r}}$), and a Visual Detail Extractor ($\mathcal{T}_{\text{v}}$). The policy network π_θ performs a single tool-selection step to determine a subset $C \subseteq \mathcal{T}$ based on the initial visual parsing and the specific causal gap being evaluated.

Stage 2: Causal Bridging & Tool Augmentation. The second stage executes the selected tools once and integrates their returned evidence into the reasoning context before producing the final prediction. The tool outputs are not merely concatenated; they enhance the agent's spatio-temporal understanding: $\mathcal{T}_{\text{s}}$ clarifies action sequences, $\mathcal{T}_{\text{f}}$ recovers transient actions, and $\mathcal{T}_{\text{r}}/\mathcal{T}_{\text{v}}$ resolve fine-grained ambiguities (e.g., hand-object contact). The final prediction y is then inferred by conditioning the model π_θ on these augmented inputs, defined by:

$$p(y|V \oplus V_{\text{aug}}, Q \oplus E_{\text{aug}}) = \pi_\theta(V \oplus V_{\text{aug}}, Q \oplus E_{\text{aug}}), \tag{1}$$

where $\oplus$ denotes modality-specific augmentation, V_{aug} and E_{aug} respectively represent augmented visual and textual evidence, while π_θ is the policy network of MLLM parameterized by weights θ.

Tool Library. As shown in Fig. 3, we integrate four diagnostic tools for causal-transition reasoning. ❶ **State Transition Tracker**: to resolve ambiguities in action order and state changes, this tool re-analyzes the video and structures the observed event chain and final state into a clear textual sequence. ❷ **Focused Evidence Retriever**: to address the omission of actions caused by uniform sampling, this tool performs dense frame extraction within a specified temporal window. ❸ **Region**

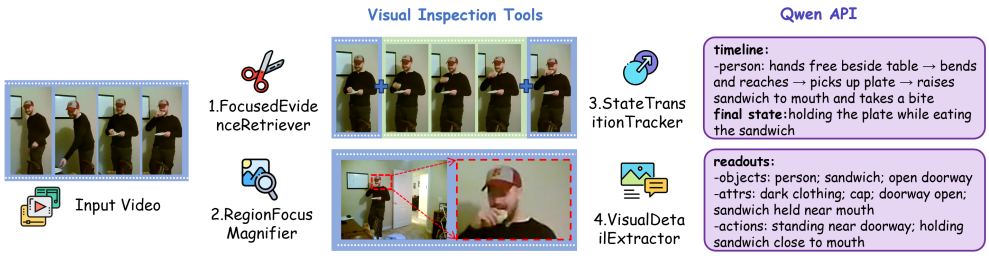

Figure 3: **Diagnostic tool library of VepAgent.** FRE retrieves dense visual evidence from selected temporal windows, while RFM crops and magnifies task-relevant spatial regions. STT summarizes observed state trajectories and the terminal state, whereas VDE extracts fine-grained visible facts as textual evidence. All tools operate exclusively on the observed portion of the video.

Focus Magnifier: to resolve local detail ambiguity (e.g., whether a hand has released an object), this tool crops and magnifies key spatial regions for focused inspection. ❹ **Visual Detail Extractor**: to mitigate overlooked fine-grained facts, this tool translates specific visual details (e.g., spatial relations) from the final frames into explicit textual descriptions, anchoring the causal reasoning in verifiable visual evidence. Details are provided in Appendix C.

3.2 Training Data Construction

A fundamental challenge when applying existing MLLMs to VEP is the absence of causal-transition reasoning chains: standard pipelines provide retrospective summarization of observed video content but lack intermediate reasoning steps—**what** the terminal observed state is, **which** action is ongoing but unfinished, and **how** the unobserved causal gap is bridged to project the future event. We therefore synthesize structured reasoning chains to construct *futurebench-4K*, providing explicit examples of these steps.

Data Collection. Our data generation begins by leveraging the Qwen3-VL-32B (Bai et al., 2025a) to explore the base VEP training datasets. For each partially observed video, we perform 8 independent policy rollouts to generate diverse causal reasoning pathways. An instance is solvable if at least one rollout matches the ground-truth under the same matching protocol used for reward computation, ensuring the prediction is derived from actual visual states rather than language-prior shortcuts.

Data Assessment. To ensure the quality and reliability of the synthesized data, we employ the Qwen-VL-Max model as an expert validator to review and score each reasoning chain. Chains containing factual errors regarding observed states, logical inconsistencies in bridging the causal gap, or language-prior shortcut behaviors are discarded. This validation yields approximately 4,000 reasoning chains for SFT. The remaining successful trajectories, confirmed as valid but not selected for SFT, are used as the training data for the RL stage. To prevent data leakage, *futurebench-4K* and all RL trajectories are constructed exclusively from the VEP training splits.

Generating Pipeline. Our generating pipeline implements a three-step causal-transition process. First, the agent identifies the terminal observed state S_{final} and the ongoing, unfinished action A_{ongoing}. Second, if critical transitions are ambiguous, it conditionally introduces diagnostic tools (e.g., State Transition Tracker or Focused Evidence Retriever) to recover missing temporal evidence. Finally, the agent bridges the *Causal Gap* by deducing how the unfinished action progresses to the next physically plausible state and the future-event prediction.

3.3 Agentic Supervised Fine-Tuning

After constructing the training data, we introduce a **cold-start phase** that prioritizes supervised fine-tuning (SFT) before applying reinforcement learning (RL). Inspired by R1-Zero (Guo et al., 2025), we initially attempt direct RL optimization to train our method. However, preliminary experiments reveal a progressive decline in the frequency of tool invocation during policy rollouts. This behavior likely arises from a distributional discrepancy in target domain between tool-enhanced visual features and model's pretraining data. Therefore, we opt to perform SFT before RL.

Specifically, we formalize each training instance as $\mathcal{W} = (\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{O})$, where $\mathcal{X}$ denotes the input modality, $\mathcal{Y}$ represents task instructions, $\mathcal{Z}$ captures the structured reasoning process, and $\mathcal{O}$ is the final target output. During SFT, the reasoning process and final answer are concatenated into a complete assistant response $\tilde{\mathcal{Z}} = [\mathcal{Z}; \mathcal{O}] = \{\tilde{z}_1, \ldots, \tilde{z}_L\}$. We optimize the model with the standard autoregressive negative log-likelihood over the entire assistant response:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\mathcal{W}\sim\mathcal{D}}\left[\sum_{t=1}^{L} \log p_\theta(\tilde{z}_t \mid \mathcal{X}, \mathcal{Y}, \tilde{z}_{<t})\right], \tag{2}$$

where $p_\theta(\cdot)$ denotes the conditional probability distribution. Thus, both the structured reasoning tokens and the final `<answer>` output are jointly optimized during SFT.

3.4 Agentic Reinforcement Learning

Group Relative Policy Optimization (GRPO). We adopt the GRPO algorithm (Shao et al., 2024) for policy optimization. GRPO employs a groupwise comparison framework to evaluate candidate responses. Specifically, for each query q paired with its ground-truth solution a from dataset D, the algorithm generates a set of rollout trajectories $\{o_1, o_2, \ldots, o_G\}$ based on the previous policy $\pi_{\theta_{\text{old}}}$. The policy π_θ is then refined through the optimization of this objective function:

$$\begin{aligned}\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G} \\ \left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}(\pi_\theta\|\pi_{ref})\right)\Bigg],\end{aligned} \tag{3}$$

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{4}$$

where β is adopted to balance the trade-off between exploration and stability during optimization. Then the advantage estimator A_i is calculated using normalized rewards from the trajectory group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{5}$$

Each trajectory o_i receives a composite reward through rule-based verification designed as follows.

Reward Design. Effective reward functions should balance prediction accuracy, causal-logic bridging, and robustness against language-prior shortcuts. To this end, we design a composite reward that integrates three components: an accuracy reward R_{acc}, a causal gap reward R_{cau}, and an anti-prior reward R_{ant}, all gated by a format reward R_{format}. The format reward serves as a hard gate to ensure structural coherence, where $R_{\text{format}}(\tau) = 1$ if the trajectory τ strictly follows the required reasoning format (e.g., containing valid **<*thinking*>** and **<*answer*>** tags), and 0 otherwise. The accuracy reward R_{acc} is a binary metric quantifying the correctness of the final prediction, where $R_{\text{acc}}(\tau) = 1$ if the predicted future event semantically matches the ground-truth event, and 0 otherwise.

Differently, the causal gap reward R_{cau} quantifies whether the model's reasoning genuinely bridges the unobserved intermediate state transition rather than directly guessing the future. Specifically, it evaluates the consistency of the predicted terminal state and causal transition with reference reasoning, while assigning greater emphasis to the transition toward the future event. The resulting process reward is further conditioned on prediction correctness to prevent incorrect predictions from receiving excessive credit.

Furthermore, the anti-prior reward R_{ant} is designed to mitigate the language-prior shortcut, where models merely select the option with the highest textual similarity to the query Q. For a given query and the correct option among K candidates, we compute the textual similarity rank k of the correct option with respect to Q using sentence-embedding cosine similarity. If the model predicts correctly, we assign a higher reward when the correct option is less textually guessable (i.e., a larger rank k). The anti-prior reward is formally defined as:

$$R_{\text{ant}}(\tau) = \begin{cases} \frac{k-1}{K-1}, & \text{if } R_{\text{acc}}(\tau) = 1 \text{ and } K > 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Table 1: Event Prediction Acc. (%) on FutureBench. best and second best results are highlighted.

Method	Frames	1-Hop	2-Hop	3-Hop	Interp.	AVG
GLM-4.1V-9B	32	29.90	41.90	52.20	47.30	44.38
LLaVA-NeXT-Video	32	48.80	49.30	40.00	44.40	45.24
Kimi-VL-A3B	32	44.30	42.80	51.30	51.90	48.87
MiMo-VL-7B	32	59.00	59.60	50.50	43.80	50.45
InternVL3-8B	32	54.30	58.00	63.20	54.40	56.72
Qwen2.5-VL-32B	32	66.50	62.70	63.20	55.20	59.94
Qwen3-VL-30B-A3B	32	65.30	70.50	76.10	62.20	66.86
Qwen3-VL-4B	32	54.91	57.51	59.20	61.15	59.09
VepAgent-4B (Ours)	**32**	**82.08**	**88.60**	**80.10**	**78.94**	**81.44**

Table 2: Event Prediction Acc. (%) on NEPBench. best and second best results are highlighted.

Method	Overall	Charades	VidSitu	YouCook2
GLM-4.1V-9B	43.73	45.98	28.18	64.48
InternVL3-8B	43.60	37.01	40.55	52.26
Kimi-VL-A3B	41.20	28.66	53.06	32.44
LLaVA-Video-7B	29.60	25.67	24.59	39.32
MiMo-VL-7B	44.53	40.47	40.83	52.46
Qwen2.5-VL-72B	47.50	35.59	41.05	64.48
Qwen2.5-VL-7B	35.23	38.90	27.32	44.15
Qwen3-VL-4B	42.17	27.40	48.96	42.09
VepAgent-4B (Ours)	**70.20**	**83.46**	**54.78**	**83.57**

This continuous formulation yields 0 for the most obvious textual guess and 1 for the least obvious, incentivizing the agent to rely on visual evidence rather than superficial textual priors. Ultimately, final reward for a trajectory τ is computed by combining these components, defined as:

$$R(\tau) = R_{\text{format}}(\tau) \cdot (0.65R_{\text{acc}}(\tau) + 0.20R_{\text{cau}}(\tau) + 0.15R_{\text{ant}}(\tau)) . \tag{7}$$

The composite reward balances prediction accuracy with causal coherence and an anti-prior penalty, encouraging visually grounded predictions.

4 Experiment

Dataset and Metrics. We evaluate our method on two public VEP benchmarks: FutureBench (Wang et al., 2026) and NEPBench (Lei et al., 2020). **FutureBench** measures the overall event prediction accuracy across different reasoning depths. Following prior work, we report accuracy on 1-Hop, 2-Hop, and 3-Hop predictions, as well as Interp. (Interpolation) and the overall AVG. **NEPBench** provides a comprehensive evaluation of Next Event Prediction capabilities across diverse real-world scenarios, containing subsets derived from Charades, VidSitu, and YouCook2. We report the Overall average accuracy and the individual accuracy on each subset. For both datasets, a prediction is considered correct only if it semantically matches the ground-truth future event.

Implementation Details.

We conduct experiments on the Qwen3-VL-4B-Instruct model. In stage one, we select approximately 4,000 training samples constructed from Sec. 3.2 to form the *futurebench-4K* dataset for SFT. For subsequent RL optimization, we use approximately 5,000 validated training samples. The training was conducted on eight NVIDIA RTX A5000 GPUs (24 GB memory each). SFT is performed for 1 epoch with an effective batch size of 16 and a learning rate of 5×10^{-6}. RL is optimized for 250 steps with 8 rollouts per sample and a learning rate of 1×10^{-6}. For component-wise ablation studies, we use $G = 4$ unless otherwise specified. More details are shown in Appendix A.1.

4.1 Comparison with Reported Baselines

FutureBench. In Tab. 1, we evaluate our method on FutureBench under standard protocols against the MLLMs reported in that table. Large-scale models such as Qwen3-VL-30B-A3B Bai et al. (2025a) and Qwen2.5-VL-32B Bai et al. (2025b) remain competitive, yet they primarily

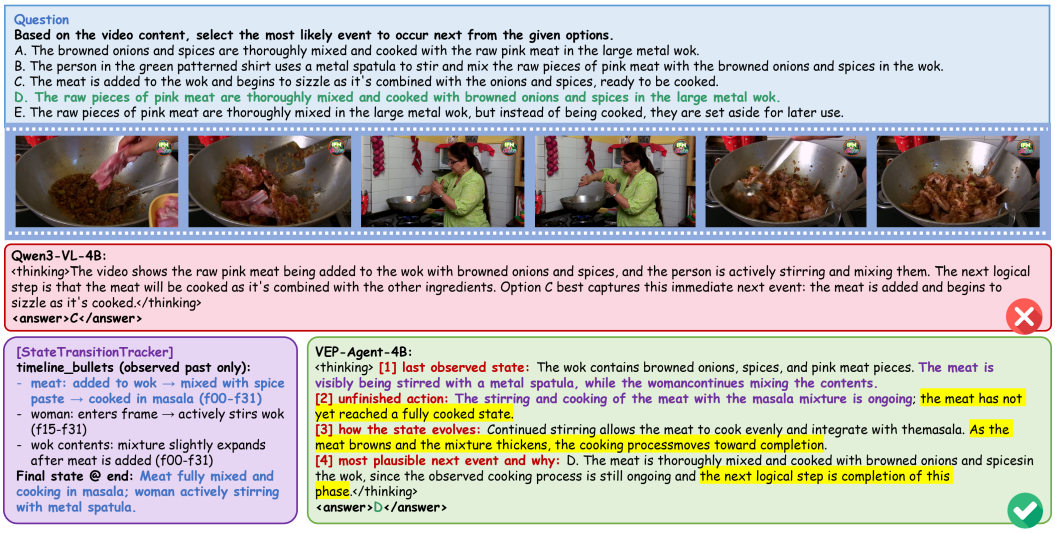

Figure 4: **Case study between Qwen3-VL-4B and VepAgent-4B.** The baseline predicts an already observed meat-addition event. StateTransitionTracker reconstructs the observed progression from meat addition to mixing and active cooking. Grounded in this evidence, VepAgent identifies the unfinished cooking process, reasons over its subsequent state evolution, and predicts the correct future event.

rely on retrospective summarization of historical observations. This passive observation often induces language-prior shortcuts, where models project future trajectories from superficial textual co-occurrences rather than causal grounding, limiting their ability to bridge unobserved intermediate state transitions.

VepAgent (4B) raises AVG accuracy by **14.58** percentage points over the prior strongest reported baseline Qwen3-VL-30B-A3B, from 66.86% to **81.44%**, under tool-augmented RL and the same 32-frame budget. Relative to the same-size Qwen3-VL-4B backbone (59.09% AVG), this corresponds to a gain of 22.35 percentage points. This gain is consistent with improved causal-transition reasoning and visual-evidence recovery.

NEPBench. In Tab. 2, we further evaluate on NEPBench. Models such as Qwen2.5-VL-72B remain competitive on specific subsets, yet they struggle with fine-grained detail ambiguity and key-frame omission in cross-domain scenarios under uniform spatio-temporal sampling. By dynamically integrating diagnostic tools to recover missing transient actions, VepAgent raises Overall accuracy by **22.70** percentage points over the prior strongest reported baseline Qwen2.5-VL-72B, from 47.50% to **70.20%**. Relative to the same-size Qwen3-VL-4B backbone (42.17% Overall), the gain is 28.03 percentage points. On the reported subsets, VepAgent also attains the highest accuracy: Charades 83.46% (next, GLM-4.1V-9B at 45.98%), VidSitu 54.78% (next, Kimi-VL-A3B at 53.06%), and YouCook2 83.57% (next, 64.48% for both GLM-4.1V-9B and Qwen2.5-VL-72B), under the same evaluation protocol.

Qualitative Analysis. Beyond the quantitative gains, Fig. 4 illustrates how tool-grounded causal reasoning corrects temporal-state misinterpretation. Although the base model recognizes the overall cooking context, it predicts an event that has already occurred. By explicitly anchoring the terminal observed state, VepAgent reasons forward from the unfinished process rather than replaying the observed action. More qualitative examples are provided in Appendix A.4.

4.2 Ablation Studies

Training Stages. In rows *(a)-(d)* of Tab. 3, we compare zero-shot, SFT-only, SFT+RL, and the full model training paradigms, respectively. The zero-shot baseline in row *(a)* yields the lowest AVG accuracy of 59.09%. SFT-only training in row *(b)* raises AVG accuracy from 59.09% to 71.59%. Combining SFT with RL in row *(c)* further raises AVG accuracy to 75.09%, and the full tool-

Table 3: Ablation studies (%) on FutureBench. Component-wise ablations are conducted with $G = 4$. best and second best results are highlighted.

Method	1-Hop ↑	3-Hop ↑	AVG ↑	Tools	1-Hop ↑	3-Hop ↑	AVG ↑
(a) Qwen3-VL-4B	54.91	59.20	59.09	*(e)* w/o. STT	82.08	71.64	75.66
(b) 4B + SFT only	77.46	**75.62**	71.59	*(f)* w/o. FRE	82.08	**75.62**	76.99
(c) 4B + SFT + RL	**79.77**	71.14	**75.09**	*(g)* w/o. RFM	**82.66**	**76.12**	**78.31**
(d) 4B + SFT + RL + Tools ($G = 4$)	**83.82**	**76.12**	**78.79**	*(h)* w/o. VDE	**83.24**	75.12	**77.46**
Reward Term	**1-Hop ↑**	**3-Hop ↑**	**AVG ↑**	**Num of Gen.**	**1-Hop ↑**	**3-Hop ↑**	**AVG ↑**
(i) $R = R_{\text{acc}}$	78.61	**76.12**	74.62	*(m) Num = 2*	**84.39**	**77.61**	78.60
(j) $R = R_{\text{acc}} + R_{\text{cau}}$	**82.08**	**78.11**	**76.42**	*(n) Num = 4*	**83.82**	76.12	78.79
(k) $R = R_{\text{acc}} + R_{\text{ant}}$	76.88	**78.11**	76.33	*(o) Num = 6*	82.08	76.72	**78.98**
(l) $R = R_{\text{acc}} + R_{\text{cau}} + R_{\text{ant}}$ (Ours)	**83.82**	**76.12**	**78.79**	*(p) Num = 8*	82.08	**80.10**	**81.44**

augmented model in row *(d)* reaches 78.79%. We attribute this progression to the difficulty of RL in directly navigating the high-dimensional causal reasoning space without prior SFT guidance. These results support our two-stage strategy, where SFT provides a structured reasoning foundation before RL optimization.

Tool-Usage. In rows *(e)*-*(h)* of Tab. 3, we show the effect of removing each specific diagnostic tool under the $G = 4$ setting. Relative to the full model in row *(d)* (78.79% AVG), removing STT in row *(e)*, FRE in row *(f)*, RFM in row *(g)*, and VDE in row *(h)* reduces AVG accuracy to 75.66%, 76.99%, 78.31%, and 77.46%, respectively, corresponding to absolute drops of 3.13, 1.80, 0.48, and 1.33 percentage points. Removing STT causes the largest drop, while the other tools provide complementary gains for future extrapolation. The full model under this setting achieves the highest AVG score, demonstrating the complementary contributions of the diagnostic tools to video event prediction.

Reward Functions. In rows *(i)*-*(l)* of Tab. 3, we analyze the effect of different reward configurations in RL training. Using only the accuracy reward *(i)* yields the lowest AVG accuracy of 74.62%, showing that accuracy and format rewards alone are insufficient to guide causal reasoning. Including the causal gap reward in row *(j)* raises AVG accuracy from 74.62% to 76.42%, illustrating that explicit causal-logic bridging provides useful logical cues even without penalizing language shortcuts. Including the anti-prior reward *(k)* similarly raises AVG accuracy from 74.62% to 76.33%, confirming that mitigating language-prior shortcuts encourages visual grounding. The full reward in row *(l)* reaches the highest AVG accuracy among these reward settings, raising it from 74.62% under the accuracy-only reward *(i)* to **78.79%**.

Number of Generations. In rows *(m)*-*(p)* of Tab. 3, we adopt different numbers of generations during the RL stage (GRPO). Increasing the number of candidate responses from 2 to 8 raises AVG accuracy from 78.60% to 81.44%, although individual metrics show some fluctuations: *Num = 2* yields the highest 1-Hop score (84.39%), while *Num = 8* (row *(p)*) reaches the highest 3-Hop and AVG scores (**80.10%** and **81.44%**). We therefore adopt *Num = 8* for our final model.

5 Conclusion

We introduced **VepAgent**, an agentic framework for Video Event Prediction (VEP) that unifies causal-transition reasoning with multimodal tool augmentation. On FutureBench and NEPBench, VepAgent (4B) raises average accuracy by 14.58 and 22.70 percentage points over the prior strongest reported MLLM baselines, from 66.86% to 81.44% and from 47.50% to 70.20%, respectively, and also improves over the same-size Qwen3-VL-4B backbone under identical protocols. The method comprises three components: (i) the *futurebench-4K* CoT dataset for supervised fine-tuning (SFT), which structures the deduction of unobserved intermediate states; (ii) a diagnostic tool library integrating state tracking, frame retrieval, and region magnification to recover missing spatio-temporal evidence and resolve visual ambiguities; and (iii) a composite reward that jointly optimizes prediction accuracy, causal coherence, and anti-prior robustness, encouraging visual grounding rather than superficial textual similarities. The present evaluation covers FutureBench and NEPBench. Broader video understanding settings remain untested. Future work will expand the diagnostic tool library and apply causal-transition reasoning to those settings.

AI USE STATEMENT

Generative AI was used to synthesize CoT reasoning trajectories for constructing training data, and to provide external textual descriptions of visual content via tool-augmented inference. AI tools also assisted with code modification, language editing, and manuscript assessment and revision. It is important to note that AI was not involved in the core research ideas, methodology, or experimental design. The authors take full responsibility for the scientific claims, experimental results, citations, and final manuscript.

REFERENCES

Gabriele Abbate, Alessandro Giusti, Viktor Schmuck, Oya Celiktutan, and Antonio Paolillo. Self-supervised prediction of the intention to interact with a service robot. *Robotics and Autonomous Systems*, 171:104568, 2024.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint arXiv:2503.24376*, 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.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pp. 753–771. Springer, 2018.

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. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026a.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5432–5443, 2026b.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13485–13495. IEEE, 2021.

Mark Granroth-Wilding and Stephen Clark. What happens next? event prediction using a compositional neural network model. In Dale Schuurmans and Michael P. Wellman (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 2727–2733. AAAI Press, 2016. doi: 10.1609/AAAI.V30I1.10344. URL https://doi.org/10.1609/aaai.v30i1.10344.

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.

Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. In *European Conference on Computer Vision*, pp. 767–785. Springer, 2020.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27992–28002. IEEE, 2024.

Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9590–9600, 2022.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video-lmm? complex video reasoning and robustness evaluation suite for video-lmms. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3642–3651. IEEE, 2025.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 8769–8784, 2020.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. MMedAgent: Learning to use medical tools with multi-modal agent. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.510. URL https://aclanthology.org/2024.findings-emnlp.510/.

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. Videvent: A large dataset for understanding dynamic evolution of events in videos. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 5128–5136. AAAI Press, 2025. doi: 10.1609/AAAI.V39I5.32544. URL https://doi.org/10.1609/aaai.v39i5.32544.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. LLaVA-Plus: Learning to use tools for creating multimodal agents, 2023. URL https://arxiv.org/abs/2311.05437.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, et al. Taco: Learning multi-modal action models with synthetic chains-of-thought-and-action. *arXiv preprint arXiv:2412.05479*, 2024.

Petraq Nako and Adam Jatowt. Navigating tomorrow: Reliably assessing large language models performance on future event prediction. *arXiv preprint arXiv:2501.05925*, 2025.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based RL. *arXiv preprint arXiv:2503.07536*, 2025.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5585–5596. IEEE, 2021.

Klaus Schmidt, Chengyu Wang, Ji-Hoon Park, Lin Wang, Santiago Munoz, et al. Video event reasoning and prediction by fusing world knowledge from llms with vision foundation models. *arXiv preprint arXiv:2507.05822*, 2025.

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.

Qile Su, Shoutai Zhu, Shuai Zhang, Baoyu Liang, and Chao Tong. Eventformer: A node-graph hierarchical attention transformer for action-centric video event prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4698–4707, 2025.

Qile Su, Jing Tang, Rui Chen, Lei Sun, and Xiangxiang Chu. Video-coe: Reinforcing video event prediction via chain of events. *arXiv preprint arXiv:2603.14935*, 2026.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. In *International Conference on Learning Representations*, volume 2026, pp. 31524–31570, 2026.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.

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. *arXiv preprint arXiv:2508.04416*, 2025.

Shitian Zhao, Haoquan Zhang, Shaoheng Lin, Ming Li, Qilong Wu, Kaipeng Zhang, and Chen Wei. PyVision: Agentic vision with dynamic tooling. *arXiv preprint arXiv:2507.07998*, 2025.

APPENDIX

In this appendix, we provide more experimental details, related work, tool library, and discussions for a comprehensive evaluation and understanding of our method. Detailed contents are as follows:

A Experiment **13**

A.1 More Experimental Details 13

A.2 Reward Implementation Details 13

A.3 More Ablation Studies 14

A.4 More Case Studies 15

B Related Work **15**

B.1 Video Event Prediction 15

B.2 Multimodal LLMs Reasoning 16

B.3 Tool-Augmented Agentic Systems 17

C Tool Library Details **17**

C.1 State Transition Tracker 18

C.2 Focused Evidence Retriever 18

C.3 Region Focus Magnifier 18

C.4 Visual Detail Extractor 18

D Prompts **19**

A EXPERIMENT

A.1 MORE EXPERIMENTAL DETAILS

Datasets. We conduct experiments on FutureBench (Wang et al., 2026) and NEPBench (Lei et al., 2020). FutureBench evaluates future-event prediction under different reasoning depths, including 1-Hop, 2-Hop, 3-Hop, and Interpolation settings. NEPBench contains diverse real-world scenarios from Charades, VidSitu, and YouCook2. All SFT and RL training trajectories are constructed exclusively from the training splits to avoid data leakage.

Evaluation metrics. We follow the evaluation protocols of the corresponding benchmarks and report prediction accuracy. For FutureBench, we report 1-Hop, 2-Hop, 3-Hop, Interp., and their overall average (AVG). For NEPBench, we report the overall accuracy together with the results on Charades, VidSitu, and YouCook2.

More Hyperparameter Configurations. We use Qwen3-VL-4B-Instruct as the backbone. SFT is performed on approximately 4K constructed CoT trajectories for one epoch with an effective batch size of 16 and a learning rate of 5×10^{-6}. For RL, we use approximately 5K validated trajectories and optimize the model with GRPO for 250 steps with a learning rate of 1×10^{-6} and a group size of $G = 8$. We set the KL coefficient β to 0.04 and the clipping ratios ϵ_{low} and ϵ_{high} to 0.2 and 0.3. All experiments are conducted on eight NVIDIA RTX A5000 GPUs with 24 GB memory.

A.2 REWARD IMPLEMENTATION DETAILS

Causal Gap Reward. For the causal gap reward, we separately evaluate the terminal observed state and the causal transition toward the future event. Let s_1 denote the generated terminal-state description in reasoning step [1], and let s_{34} denote the concatenation of reasoning steps [3] and [4]. We compare them with the corresponding reference terminal state r_1 and causal-transition reference r_4, respectively. Semantic similarity is computed using MiniLM sentence embeddings with cosine

similarity. The raw causal consistency score is defined as

$$g_{\text{raw}} = \frac{0.45\,\text{sim}(s_1, r_1) + 0.55\,\text{sim}(s_{34}, r_4)}{0.45 + 0.55}. \tag{8}$$

If only one reasoning block is available, we compute the score using the available block with its weight renormalized. If both blocks are missing, the complete `<thinking>` content is used as a fallback.

To prevent an incorrect prediction from receiving excessive process reward, the causal reward is further conditioned on prediction correctness:

$$R_{\text{cau}} = \begin{cases} \min(1,\, 0.20 + 0.80 g_{\text{raw}}), & R_{\text{acc}} = 1, \\ \min(0.35,\, 0.20 g_{\text{raw}}), & R_{\text{acc}} = 0. \end{cases} \tag{9}$$

Anti-Prior Reward. For the anti-prior reward, we encode the query and each candidate option using a MiniLM sentence encoder and compute their cosine similarities. Candidate options are ranked according to their textual similarity to the query, and the rank of the correct option is used to compute R_{ant} as defined in the main paper.

Reward Ablations. For reward-component ablations, the format reward remains active as the hard structural gate. When removing an individual reward term, the original weights of the remaining terms are retained without renormalization.

A.3 More Ablation Studies

To further validate our design choices, we conduct comprehensive ablation studies on the FutureBench dataset under the same G=4 setting as the component-wise ablations in the main paper.

Table 4: Ablation on tool selection strategies.

Strategy	1-Hop	3-Hop	AVG	Avg Calls
(a) No Tools	79.77	71.14	75.09	0.00
(b) All Tools (Static)	76.30	70.65	74.91	4.00
(c) Random Tools	82.08	77.61	76.99	1.85
(d) Dynamic (Ours)	**83.82**	**76.12**	**78.79**	1.13

(a) Tool Selection Strategy. In Tab. 4, we compare our dynamic selection against: (a) No Tools, (b) All Tools (statically invoking all 4 tools), and (c) Random Tools. Static tool invocation in (b) yields 74.91% AVG, slightly below the tool-free baseline (75.09%), as forcibly injecting all tool outputs can introduce irrelevant evidence that distracts the model from the true causal cues. Random selection in (c) raises AVG accuracy from 75.09% to 76.99% over No Tools, but remains below Dynamic and uses more average tool calls (1.85 vs. 1.13), reflecting the cost of sometimes retrieving uncorrelated spatial or temporal evidence. Our Dynamic Selection in (d) achieves the highest AVG accuracy (78.79%) and the best 1-Hop score (83.82%), with an average call count of 1.13. Random attains a higher 3-Hop score (77.61% vs. 76.12%), yet Dynamic remains strongest on AVG. This demonstrates that the RL agent learns to retrieve evidence selectively when the confidence of S_{final} and A_{ongoing} is low. Relative to Random, Dynamic raises AVG accuracy from 76.99% to 78.79% while using fewer tool calls (1.13 vs. 1.85).

Table 5: Ablation on tool synergy between temporal and spatial dimensions.

Tool Configuration	1-Hop	3-Hop	AVG
(a) Temporal Only (STT + FRE)	82.08	75.62	77.84
(b) Spatial Only (RFM + VDE)	80.92	71.64	75.38
(c) All (Ours)	**83.82**	**76.12**	**78.79**

(b) Synergy of Spatio-Temporal Tools. In Tab. 5, We evaluate the agent using only temporal or spatial tools. Relying solely on spatial tools (RFM, VDE) in (b) allows the model to verify local details but fails to capture chronological dynamics, yielding the weakest 3-Hop accuracy (71.64%) and the lowest AVG (75.38%). Conversely, temporal tools (STT, FRE) in (a) clarify action sequences and reach 77.84% AVG, but still fall short of the full model on 1-Hop and AVG. Combining both dimensions in (c) achieves the best overall results (83.82% / 76.12% / 78.79% on 1-Hop / 3-Hop /

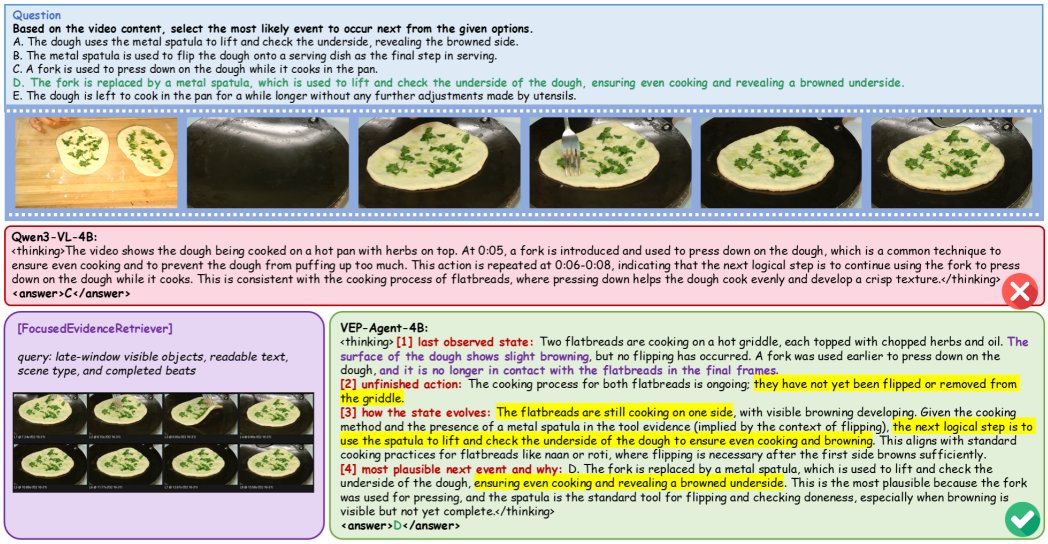

Figure 5: **Additional case study with FocusedEvidenceRetriever.** The baseline repeatedly predicts the earlier fork-pressing action. FRE retrieves dense late-window visual evidence showing that the fork interaction has ended while the flatbread continues browning. VepAgent uses this temporal evidence to reason toward the subsequent cooking transition.

AVG), indicating that temporal tracking and spatial magnification are complementary for resolving visual-evidence omissions in VEP. Specifically, temporal tools narrow down the critical time windows, which subsequently allows spatial tools to perform precise verifications on the exact frames, forming a cohesive evidence-gathering pipeline.

A.4 More Case Studies

To further illustrate how diagnostic tools support causal-transition reasoning, we provide two additional qualitative examples. These cases highlight two representative failure modes in VEP: missing temporal evidence and fine-grained state ambiguity. In both cases, the baseline tends to repeat an event that has already occurred, whereas VepAgent uses additional evidence from the observed video to better determine the terminal state and reason toward the subsequent event.

Recovering Missing Temporal Evidence. As shown in Fig. 5, sparse observation can make the model overemphasize an earlier salient action and incorrectly treat it as the future event. By retrieving denser evidence from the relevant late temporal window, FRE reveals that the fork interaction has already finished while the cooking process remains ongoing. This additional evidence helps VepAgent distinguish the completed action from the unfinished process and therefore reason toward the next plausible cooking event.

Resolving Fine-Grained State Ambiguity. Fig. 6 demonstrates a complementary failure mode in which the overall scene is correctly recognized but the completion state of an action is misinterpreted. VDE extracts explicit visual facts from the final observed frames, allowing VepAgent to verify that the onion-addition stage has already completed and that the cooking process has progressed further. Grounded in this terminal-state evidence, the model avoids repeating the observed event and instead predicts the subsequent ingredient-related transition.

B Related Work

B.1 Video Event Prediction

Video Event Prediction (VEP) is closely related to research on script event prediction, action anticipation, and temporal video understanding. Early studies on script event induction modeled depen-

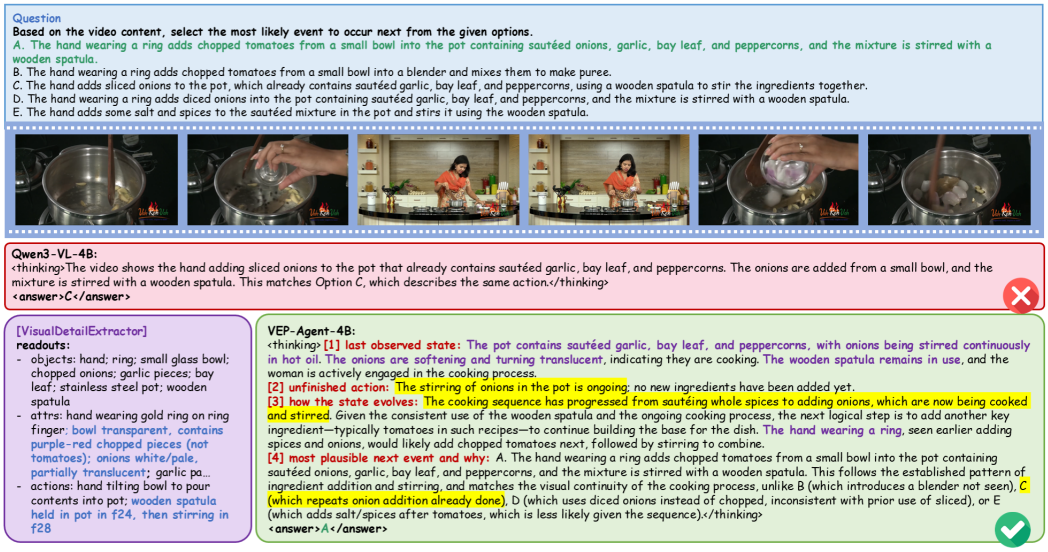

Figure 6: **Additional case study with VisualDetailExtractor.** The baseline repeats the already observed onion-addition event. VDE verifies fine-grained visual states, including partially translucent onions being stirred in the pot. VepAgent therefore recognizes that onion addition has completed and reasons toward the next ingredient stage.

dencies among events in textual narratives to infer plausible subsequent events (Granroth-Wilding & Clark, 2016). In the visual domain, action anticipation extended this forward-looking objective to videos by predicting future actions from partial observations (Girdhar & Grauman, 2021). Large-scale video datasets such as EPIC-Kitchens (Damen et al., 2018) and VidSitu (Sadhu et al., 2021) further provided structured annotations of actions, objects, and semantic roles, facilitating the modeling of dynamic event evolution in complex real-world videos.

More recent studies have moved from action-level anticipation toward explicit future-event reasoning. VLEP (Lei et al., 2020) formulates next-event prediction from partially observed videos, requiring models to identify the event most likely to occur after the observed context. To capture richer temporal and semantic structures, VidEvent (Liang et al., 2025) studies structured video events, while EventFormer (Su et al., 2025) models hierarchical dependencies among events for action-centric prediction. With the rapid development of multimodal large language models, FutureBench (Wang et al., 2026) further evaluates future-event reasoning across different temporal reasoning depths. Video-CoE (Su et al., 2026) introduces a Chain-of-Events paradigm to explicitly organize observed temporal events and strengthen the logical connection between video history and future outcomes.

Despite these advances, existing approaches mainly improve temporal representation, event structure modeling, or event-chain reasoning based on the available observations. They may still struggle when the terminal observed state is ambiguous, when an ongoing action is incomplete, or when critical evidence is omitted by sparse visual sampling. Different from these methods, VepAgent explicitly models the causal transition from the terminal observed state and unfinished action toward the future event, while further introducing diagnostic tools to recover missing temporal and fine-grained visual evidence.

B.2 Multimodal LLMs Reasoning

Recent advances in large language models have demonstrated that reinforcement-learning-based post-training can improve structured reasoning capabilities, as exemplified by OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025). This paradigm has subsequently been extended to multimodal large language models (MLLMs), where reasoning requires jointly interpreting linguistic instructions and visual observations. Representative studies explore reinforcement learning for

multimodal question answering (Peng et al., 2025), visual segmentation and grounding (Liu et al., 2025), and other vision-language reasoning tasks, demonstrating the potential of reasoning-oriented post-training beyond purely textual domains.

Video reasoning introduces additional challenges due to temporal ordering, evolving object states, and long-range dependencies. Video-R1 (Feng et al., 2026a) extends reinforcement-learning-based reasoning to video MLLMs and encourages models to better exploit temporal information. Recent work such as OneThinker (Feng et al., 2026b) further investigates unified reasoning across image and video tasks. Meanwhile, studies of video reasoning models have shown that stronger reasoning capability does not necessarily guarantee stronger visual utilization, as MLLMs may still under-utilize visual evidence or rely on language-derived priors when solving temporally complex problems (Chen et al., 2025; 2026).

These limitations are particularly critical for VEP, where subtle differences in the final observed state can lead to substantially different future trajectories. Different from generic multimodal reasoning approaches, our work focuses on causal-transition reasoning for future-event prediction. VepAgent explicitly grounds the reasoning process in observed state transitions and further combines reinforcement learning with causal-gap and anti-prior objectives to encourage visually grounded future prediction.

B.3 Tool-Augmented Agentic Systems

Recent progress in multimodal reasoning has increasingly explored the use of external tools to acquire, transform, or verify task-relevant visual evidence. Early approaches such as FAST (Sun et al., 2025) and MVoT (Li et al., 2025) incorporate visual information into intermediate reasoning, enabling multimodal chains of thought rather than relying solely on textual reasoning traces. LLaVA-Plus (Liu et al., 2023) further investigates tool-use capabilities for multimodal agents by enabling models to invoke external visual and vision-language skills. Visual Program Distillation (Hu et al., 2024) follows a complementary direction by generating and executing visual programs with specialized tools and distilling successful reasoning traces into vision-language models.

More recent work has moved toward explicitly agentic visual reasoning. TACO (Ma et al., 2024) trains multimodal action models with synthetic chains of thought and action, allowing reasoning steps to be interleaved with external tool invocation. PyVision (Zhao et al., 2025) enables MLLMs to dynamically generate and execute visual-processing programs according to the current reasoning problem, further improving the flexibility of tool-assisted perception. These approaches show that multimodal models can benefit from actively acquiring task-specific evidence rather than passively relying on a fixed visual representation.

Tool augmentation has also begun to emerge in video reasoning. VITAL (Zhang et al., 2025) introduces a tool-augmented framework for long-video reasoning, where relevant video evidence can be densely re-sampled on demand and incorporated into the subsequent reasoning process. Such methods demonstrate the effectiveness of active evidence acquisition when important visual information is sparsely distributed across long videos.

However, most existing tool-augmented frameworks are designed for general visual reasoning, multimodal question answering, or long-video understanding, rather than the specific causal requirements of future-event prediction. In contrast, VepAgent introduces a VEP-oriented diagnostic tool library that targets chronological state ambiguity, missing transient actions, local-region ambiguity, and overlooked fine-grained visual facts. The selected tools operate exclusively on the observed video and provide additional evidence for causal-transition reasoning, enabling the agent to bridge the unobserved transition toward the most plausible future event.

C Tool Library Details

This section provides a detailed specification of the multimodal diagnostic tools integrated into our VepAgent framework. These tools are designed to resolve visual-evidence omissions and spatio-temporal ambiguities during Video Event Prediction (VEP) and are categorized into chronological state tracking, temporal evidence retrieval, spatial magnification, and fine-grained fact extraction.

Table 6: Implementation details of the diagnostic tool library.

Tool	Input	Output	Failure Handling
State Transition Tracker (STT)	observed video	textual event chain & final state	fallback to global video caption
Focused Evidence Retriever (FRE)	video & temporal window	dense frame sequence (contact sheet)	ignored if temporal window is invalid
Region Focus Magnifier (RFM)	video & predicted spatial boxes	magnified image patch	ignored if boxes are invalid
Visual Detail Extractor (VDE)	final frames & scene query	textual fine-grained facts	fallback to original query

Tool Execution Protocol. For each input, the policy performs a single tool-selection step and selects a subset of diagnostic tools $C \subseteq \mathcal{T}$. The selected tools are executed once, and their returned evidence is incorporated into the reasoning context before the final causal prediction. No additional tool-selection round is performed after receiving the tool outputs. All diagnostic tools operate exclusively on the observed portion of the video and never access future frames.

Implementation Backends. FRE and RFM are implemented as lightweight visual-processing operators using OpenCV and Pillow. STT and VDE use a DashScope-hosted VLM to respectively construct textual state-transition trajectories and extract fine-grained visible facts.

C.1 State Transition Tracker

To resolve ambiguities in action order and state changes, our framework incorporates a **State Transition Tracker (STT)**. Its primary function is to chronologically re-analyze the observed video and structure the historical event chain into a clear textual sequence. This tool is dynamically invoked when the initial observation suggests conflicting state transitions (e.g., whether the hand transitioned from empty to holding, or vice versa). By explicitly mapping out the sequence of past actions and identifying the exact final observed state (e.g., "The person is still holding the cup above the table"), the agent can accurately anchor its causal-transition reasoning. This capability is critical for mitigating state/sequence misreading and directly establishes the foundational premise from which the future event is extrapolated.

C.2 Focused Evidence Retriever

To address the omission of transient actions caused by uniform frame sampling, we introduce the **Focused Evidence Retriever (FRE)**. This tool is specifically designed to capture fleeting actions that occur between sparsely sampled frames (e.g., missing the exact moment a cup is placed on a table). When the agent identifies a critical temporal gap where an ongoing action is suspected but unobserved, the FRE performs dense frame extraction within a specified temporal window. These densely sampled frames are then stitched into a contact sheet and fed back into the agent as augmented visual evidence. By recovering these missing spatio-temporal cues, the FRE ensures that crucial intermediate state transitions are not overlooked, enabling robust causal bridging.

C.3 Region Focus Magnifier

To resolve local detail ambiguity, our framework employs a **Region Focus Magnifier (RFM)**. This tool is dynamically invoked when the global view is insufficient to conclusively verify micro-state changes, such as hand-object contact or the precise orientation of a tool. By cropping and magnifying key spatial regions centered on specific human-object interactions, the RFM allows the agent to perform a focused inspection of critical details. This capability is indispensable for accurately determining whether an ongoing action has been completed or is still in progress (e.g., verifying if a hand has truly released a cup or if a door handle has been turned). By magnifying these subtle cues, the RFM directly mitigates failures arising from local detail blur.

C.4 Visual Detail Extractor

To mitigate overlooked fine-grained visual facts, we leverage the **Visual Detail Extractor (VDE)**. This component translates specific, easily ignored visual details from the final observed frames into explicit textual descriptions. When the agent needs to verify specific physical conditions to formulate a causal hypothesis, the VDE provides grounded factual statements (e.g., "The person's

right hand is still holding the cup," or "The cup is moving downward toward the table"). By explicitly textualizing these verifiable visual facts without predicting the future, the VDE anchors the agent's reasoning in concrete evidence. This strict factual grounding prevents the model from taking language-prior shortcuts, forcing it to base its causal gap bridging on actual visual states rather than superficial textual similarities between the query and options.

D PROMPTS

This section provides the prompt templates used in our agentic reasoning pipeline. As described in the main paper, VepAgent follows a two-round interaction pattern. In the first round, the model analyzes the current visual evidence and determines whether additional diagnostic evidence is needed, selecting an appropriate subset of tools when necessary. The selected tools are executed once on the observed portion of the video, and their outputs are returned to the model. In the second round, the model integrates the original observations with the retrieved evidence and performs structured causal-transition reasoning, identifying the terminal observed state, unfinished action, plausible state evolution, and final future-event prediction. The following figures present the corresponding prompt templates used in our framework.

```
# ROUND 1 — Tool Selection

You are an expert in video future-event prediction.
Use the observed video (and tool evidence if any) to verify future-event options as hypotheses, not as visual evidence.
Every reply MUST contain both <thinking>...</thinking> and <answer>...</answer>

Observed VIDEO + future-event QUESTION.
Find any prediction-critical *observed* uncertainty (ending state / event order / small region / fine detail), then choose tools to resolve it. Tools cannot reveal the future.

Tools:
1. StateTransitionTracker — Check whether one observed state comes before another (A → B). Write the two states to compare. e.g. `hand: empty -> holding cup`
2. FocusedEvidenceRetriever — Pull extra frames around a short moment that is hard to read. e.g. `near end: still holding vs already released`
3. RegionFocusMagnifier — Zoom a small blurry region so a detail is readable. e.g. `final frame, door handle: open vs closed`
4. VisualDetailExtractor — Answer a concrete visible-fact question from existing frames. e.g. `final frame: cup on table? yes/no`

Rules:
- If no real uncertainty in what is already visible, answer all tools as `no`.
- Each request: one short inspect goal on what is already visible.

## Output Format

In <answer>, answer each tool one-by-one: write the inspect goal, or `no`.

<thinking>
[1] Video situation:
2–3 sentences on actors, key actions/changes, and ongoing direction.

[2] Uncertainty -> tool:
name one observed gap, or none; pick tool(s) only if needed.
</thinking>
<answer>
StateTransitionTracker: yes-or-no
FocusedEvidenceRetriever: yes-or-no
RegionFocusMagnifier: yes-or-no
VisualDetailExtractor: yes-or-no
</answer>
```

```
# ROUND 2 — Forward Causal Future-Event Reasoning

You are an expert in video future-event prediction.
Use the observed video (and tool evidence if any) to verify future-event options as hypotheses, not as visual evidence.
Every reply MUST contain both <thinking>...</thinking> and <answer>...</answer>

Use the observed video and the tool evidence below. Reason from the video
situation to the most plausible future event, then choose a letter.

Important:
- Options are hypotheses to verify, not visual evidence already seen.
- Infer from the video first.

## Input
Question:
{question}

Answer Options:
{format_options}

Tool Evidence:
{tool_block}

Reply format:
<thinking>
Briefly cover, when useful:
last observed state -> unfinished action (or none) -> how the state evolves -> most plausible next event and why
</thinking>
<answer>A, B, C or D</answer>
```