

# FOSTERING VIDEO REASONING VIA NEXT-EVENT PREDICTION

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
Anonymous authors  
Paper under double-blind review

## ABSTRACT

Next-token prediction serves as the foundational learning task that enables reasoning in LLMs. But what should the learning task be when aiming to equip MLLMs with *temporal reasoning* capabilities over video inputs? Existing tasks such as video captioning primarily promote modality alignment, while video question answering typically relies on annotations from humans or much stronger MLLMs. To address this gap, we propose **next-event prediction (NEP)**, a learning task that harnesses future video segments as a rich, self-supervised signal to foster temporal reasoning. We segment each video into **past** and **future** frames: the MLLM takes the past frames as input and predicts events in the future, thereby encouraging the model to reason temporally in order to complete the task. To study this learning task, we curate **V1-33K**, a dataset comprising 33,000 automatically extracted videos spanning diverse real-world scenarios. Using the same videos, we further explore a range of video instruction-tuning task data to provide controlled comparisons and isolate the effect of NEP. To evaluate progress, we introduce **FutureBench** to assess coherence in predicting unseen future events. Experiments validate that NEP offers a scalable and effective training task for fostering temporal reasoning in MLLMs.

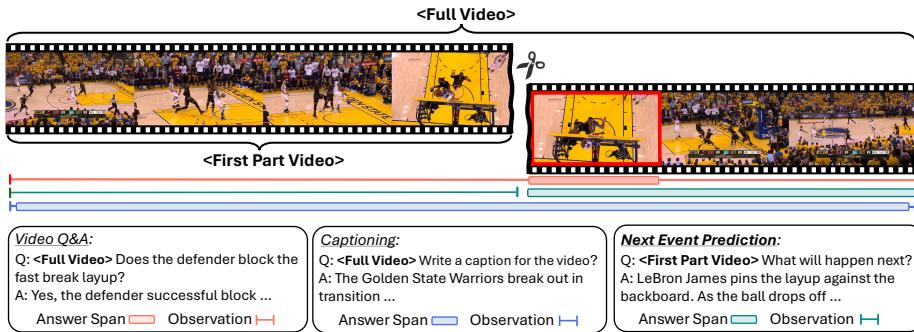
## 1 INTRODUCTION

Recent progress in multimodal large language models (MLLMs) has significantly advanced video understanding capabilities (Hurst et al., 2024; Team, 2025). Video instruction tuning typically involves *learning tasks* such as video question answering, captioning, and grounding, etc., which mainly emphasize visual perception skills such as object identification, event recognition, timestamp inference, and factual recall based on observed video frames (Bai et al., 2025; Li et al., 2024c; Lin et al., 2023; Zhang et al., 2024b). While these tasks facilitate cross-modal alignment—an essential step in integrating visual encoders with language models (Liu et al., 2023)—they often neglect the *temporal* dimension that distinguishes videos from static images. For instance, video question answering frequently relies on key frames (Cores et al., 2025), and video captioning tends to map frames to text in a frame-by-frame manner, limiting the model’s ability to understand dynamic event progression. Moreover, tasks like question answering and timestamp grounding typically require annotations by humans, raising scalability challenges. This leads to a natural question:

*What learning task should be employed to effectively equip MLLMs with temporal reasoning capabilities over video inputs?*

To bridge this gap, we propose **Next-Event Prediction (NEP)**, a learning task explicitly designed to foster temporal reasoning in MLLMs. Instead of providing the entire video as input, NEP splits each video into *past* and *future* frames. Given only the past frames, the model predicts captions that describe the future segment, as illustrated in Figure 1. NEP encourages MLLMs to reason beyond the visible scene, inferring causes, effects, and likely outcomes. The NEP naturally drives the model to integrate visual perception from the visual encoder with commonsense knowledge in the LLM.

To systematically study the effectiveness of NEP as a learning task, we construct **V1-33K**, a large-scale dataset comprising approximately 33,000 automatically curated videos. Each instance consists of an observable (past) video segment paired with a textual caption describing its subsequent continuation, which serves as the supervision target. The V1-33K is derived from the video data source used by previous works for video-instruction tuning (Zhang et al., 2024b; Li et al., 2024a), and thus naturally includes data suitable for other learning tasks, such as captioning and question answering.



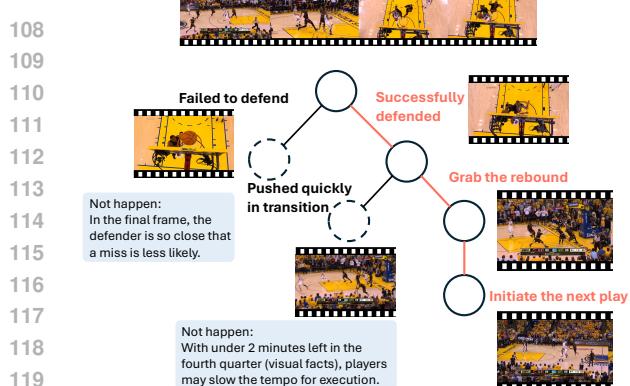


Figure 2: **Reasoning structure underlying NEP.** Each node is a potential event or action derived from visual cues, branching into alternative scenarios such as failing to defend or being pushed in transition. The red line highlights actual event sequence observed in the video. Comments provide reasoning for less likely scenarios.

future into textual descriptions of the anticipated video content. These works typically define the “*future*” at short horizons, e.g., the next frame or next action; and optimize low-level objectives, e.g., representation learning, which has proven effective for video encoder pretraining.

Our work is distinct in that it focuses on high-level, semantic future-event prediction expressed in natural language. This task requires MLLMs to perceive temporally visual information via the vision encoder and then integrate it with pre-trained commonsense knowledge in the LLM to generate open-ended natural language responses. The goal is different, therefore, rather than improving a video encoder, as previous methods attempted, we aim to improve the cross-modal projector and LLM of the MLLM while freezing visual encoder parameters. Due to the limited space, we defer the discussion of Related Work to Appendix A.

### 3 NEXT-EVENT PREDICTION

Our work focuses on the study of NEP as a learning task for MLLMs. In this section, We first formalize the NEP task, then describe the constructed dataset for studying it, and finally detail training strategies (e.g., SFT, CFT, and Distillation) used in combination with NEP.

#### 3.1 FORMULATION

We formulate NEP in a video as a sequence-to-sequence language modeling problem conditioned on video frames. Supposing  $V = [v_1, v_2, \dots, v_T]$  represents a sequence of video frames (or clips), a cut-off time  $t < T$  is chosen to split the full video into an observed part  $V_{\leq t} = [v_1, \dots, v_t]$  (*past* frames) and a future part  $V_{>t} = [v_{t+1}, \dots, v_T]$  (*future* frames). The goal is to train an MLLM that takes  $V_{\leq t}$  as input and generates a textual  $Y$  of events in  $V_{>t}$ . In practice, for SFT training strategy,  $Y$  can be simply represented by the token sequence of captions in future frames. For the details of other training strategies, like CFT and distillation, we detail in Section 3.3.

**Why NEP can be effective.** Because the target text describes events not visible in  $V_{\leq t}$ , NEP pushes the model beyond perception (e.g., object detection or action recognition) to prediction. It requires MLLMs to infer object interactions and dynamics, project them forward in time, and maintain narrative coherence. To predict plausible next events, MLLMs must integrate two information sources: (i) perceived visual evidence from the visual encoder and (ii) commonsense world knowledge, such as physics, social norms, and typical human behavior from the LLM, as reasoning drivers. NEP training therefore combines perception with knowledge-guided causal inference, resulting in improved temporal understanding and reasoning, which are required for complex video understanding.

#### 3.2 DATASET CONSTRUCTION FOR STUDYING NEP

To systematically study NEP, we build **V1-33K**, a dataset produced by a simple yet effective pipeline that converts videos into training instances for NEP.

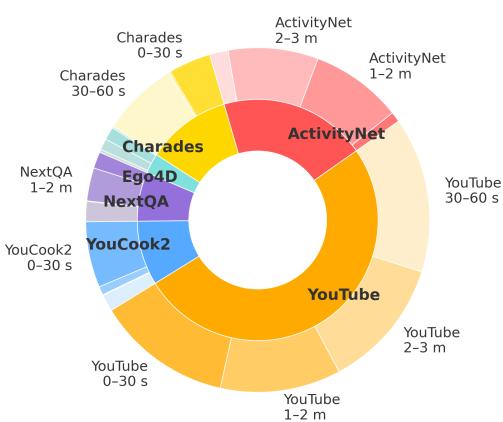


Figure 3: **Distribution of data source and video length in V1-33K.** The inner circle illustrates the distribution of data sources. The outer circle further segments each source according to video length categories. Only length categories comprising more than 4% of the dataset are labeled explicitly in the outer circle.

162 **Pipeline overview.** Given a video data source, we extract a *past* video segment (model input) and a  
 163 *future* video segment together with text supervision signal (model target) via the following procedure.  
 164

- 165 1. **(Optional) Caption Analysis.** An LLM parses the video caption to identify distinct scenes and  
 166 a causal pivot (the point where “what has happened” transitions into “what will happen”). This  
 167 yields a *text split point*.
- 168 2. **Grounding and Segmentation.** We ground the *text split point* to the video timeline using an  
 169 MLLM, obtaining a timestamp  $t$ . The video is cut at  $t$  into **past** and **future** parts, and the caption  
 170 is split accordingly into a *past-caption* and a *future-caption* (target).
- 172 3. **(Optional) Reasoning & Critique.** The past caption is processed by a *text* reasoning model to  
 173 generate predictions and reasoning traces, which are then critiqued by another LLM.

174 Note, both *Caption Analysis* and *Reasoning & Critique* are optional. Caption analysis aims to improve  
 175 the challenge of NEP by suggesting a split point in text format to avoid easy and less meaningful  
 176 next-event prediction. If analysis is skipped, we fallback to sampling a mid-sentence pivot from the  
 177 caption as the *text split point* and ground it and then split video in the same way. The reasoning and  
 178 critique step aims to rollout reasoning traces and critiques data for critique fine-tuning and distillation  
 179 in our ablation study of different training strategies. **While both *Caption Analysis* and *Reasoning &***  
 180 ***Critique* are technically optional—the pipeline can still produce future prediction data without them.**  
 181 **We enable them when constructing the V1-33K because they improve the quality and challenge level**  
 182 **of the constructed data.** We leave more details in Appendix B.

183 **NEP yields auto-labeled self-supervision signals.** NEP leverages signals already present in the  
 184 data. The only prerequisite of the signal construction (*Grounding and Segmentation*) is timestamp  
 185 grounding ability, aligning the caption split with the corresponding video cut. Following the common  
 186 training recipe (Bai et al., 2025; Chen et al., 2024b; Zhang et al., 2024b), the model undergoes  
 187 CLIP pre-training (for visual encoder), vision-language alignment (modality alignment), and then  
 188 end-to-end fine-tuning (advanced abilities). NEP asks an MLLM to perceive visual information  
 189 and then integrate it with the LLM module’s commonsense knowledge to perform reasoning and  
 190 predict future events. Considering that NEP requires MLLMs to perceive visual information and then  
 191 integrate it with the LLM module’s commonsense knowledge for reasoning and future prediction, it  
 192 is reasonable to take NEP as the learning task in the late fine-tuning stage for unlocking advanced  
 193 temporal ability, where the basic timestamp referencing/grounding ability has already been learned  
 194 in an early fine-tuning stage. With this in place, the trained MLLM can reliably perform timestamp  
 195 grounding, yielding auto-labeled self-supervision. Even considering the caption analysis, the LLM of  
 196 MLLM is sufficient to complete this lightweight task.

197 **Collection and coverage of V1-33K.** Following this pipeline, we process thousands of videos  
 198 from diverse sources (e.g., YouTube, YouCook2, NextQA, Charades, ActivityNet), containing  
 199 33,000 *past-future* pairs. The dataset spans rich scenarios—physical events (spills, collisions, object  
 200 interactions), human interactions (arguments leading to reactions, pranks leading to surprises), and  
 201 sports (setups leading to goals or failures). Source and duration statistics are summarized in Fig. 3.

202 **Data quality and source bias.** The quality of data has a strong influence on the performance of  
 203 foundation models. In our work, to study the effect of NEP as a learning-task in comparison to  
 204 others, we intentionally use the same video data source as other common instruction tuning tasks  
 205 such as video captioning and QA. This allows training the same MLLMs on the same videos but  
 206 with different tasks (captioning/QA vs. NEP), providing a controlled comparison that holds video  
 207 quality and source bias fixed. The quality of NEP data is also determined by its textual supervision.  
 208 Given the past video segment, plausible continuations vary widely, resulting in varying difficulty of  
 209 next-event prediction—some trivial, some extremely challenging, others reasonably for the currently  
 210 trained MLLMs. This mirrors next-token prediction in language, where the difficulty of the next  
 211 token also varies; its success for foundation models comes from its ease of scaling up. We follow the  
 212 same principle: our NEP data come from a simple, low-cost pipeline that automatically generates  
 213 supervision. Our method is cheaper than video question answering. Note, even for humans at the  
 214 PhD level, asking constructive questions is hard. For our data, we use the caption analysis step to  
 215 cheaply improve target quality and rely on scalable automatic generation and accept some variation  
 in difficulty. Despite the mixed quality of the data, our experimental results indicate the promise of  
 using NEP as the learning task. And we expect NEP to benefit even more as data quality improves.



Figure 4: **Task demonstration of FutureBench.** This figure presents two paradigms for future event prediction: Extrapolation and Interpolation. In the **Extrapolation task (Top)**, the model observes the initial video (Current Event) and is required to sequentially predict a series of future events (Caption 1 → Caption 2 → Caption 3 → ...) leading up to the final event (Caption N). In the **Interpolation task (Bottom)**, the model observes the initial video (Current Event) and is provided with the first future event (Caption 1), an anchor future event (Caption K), and the final event (Caption N) and must infer the most plausible intermediate events that bridge the temporal gap. Distractors involve Caption 0 of the current event to require the model to understand the given video. Questions and answer options above are simplified for clarity and brevity.

### 3.3 VIDEO INSTRUCTION-TUNING STRATEGIES

Besides the comparison in the task formulation, we further investigate four video instruction-tuning strategies on the NEP task. Each training strategy leverages specific annotations and structures from the V1-33K data pipeline, ranging from ground-truth next-event descriptions to critique and reasoning traces. We consider the encoder-decoder architecture model akin to recent MLLMs, LLaVA (Liu et al., 2023), where a vision encoder  $E$  processes the video frames and produces a sequence of visual embeddings, and a language decoder  $D$  attends to these embeddings to generate text. Specifically, for each input video  $V_{\leq t}$ ,  $E$  extracts frame features, and these features are fed into  $D$  through a cross-attention mechanism. The decoder is then prompted to output the next-event description. During training, we supervise  $D$  to match the ground truth next-event description using a standard language modeling loss, cross-entropy over the next token. We explore four distinct Video Instruction-Tuning strategies: supervised fine-tuning (SFT), critique fine-tuning (CFT), distillation tuning (Distill), and mix tuning (Mix), leveraging ground-truth video captions, critiques from GPT, and structured reasoning traces from DeepSeek. Details of tuning strategies are provided in Appendix D.

## 4 FUTUREBENCH

To advance the evaluation of MLLMs in temporal reasoning—specifically in forecasting future events from observed video—we introduce **FutureBench**, a benchmark designed to assess models’ ability to infer plausible event progressions leading to a specified outcome. This task demands both strong visual perception and commonsense reasoning. Unlike prior video Q&A benchmarks, which focus on answer extraction from visible frames (Chen et al., 2025; Xiao et al., 2021), FutureBench emphasizes temporal-causal reasoning toward achieving unobserved future goals.

We formalize the evaluation in a multiple-choice question-answering format. Each video in FutureBench is paired with a clearly defined task end goal or event outcome, termed an “*end anchor*”, which is derived from the final state of the full video. This design reflects the principle that real-world narratives typically take goal-driven paths, and it serves to limit the search space for potential future events. Given the anchor, the model must reason both *forwards* and *backwards* to deduce the plausible intermediate steps or events that lead to a specified outcome.

### 4.1 TASK SETTINGS: MULTI-HOP PREDICTION

A defining characteristic of FutureBench is its structured division into tasks with varying logical-hop distances, that is, the number of inferential steps or missing events the MLLM must predict. This design enables a comprehensive evaluation of both in-distribution performance on single-hop (1-hop) reasoning tasks and out-of-distribution generalization to more complex multi-hop reasoning involving extended event sequences. Accordingly, FutureBench is organized into two primary subtasks:

Table 1: **Performance comparison across different video instruction tuning tasks on Qwen2.5-VL-7B-Instruct.** G-Avg. and T-Avg. represent the average performances of all general and temporal benchmarks, respectively. Instruct represents the original performances without additional training.

Task	General Benchmark				Temporal Benchmark				
	VMME(w/o sub)	MVB	LVB <sub>val</sub>	G-Avg.	TB	TC	SB-R1	FB	T-Avg.
Instruct	59.8	65.3	55.9	60.3	35.4	73.8	37.1	52.6	49.7
Full Observed Video									
Captioning	<b>60.6</b>	66.2	53.2	60.0	37.0	72.2	33.6	55.8	49.7
MCQA	57.4	65.2	53.0	58.5	32.1	65.5	33.0	60.3	47.7
OEQA	59.8	<b>66.8</b>	54.6	60.4	36.6	74.0	35.4	58.8	51.2
Partially Observed Video									
NEP	60.0	66.5	<b>56.3</b>	<b>60.9</b>	<b>38.6</b>	<b>74.7</b>	<b>39.5</b>	<b>61.3</b>	<b>53.5</b>

**Future Event Prediction—Extrapolation.** The extrapolation requires the model to predict a sequence of future events that logically connect the initial observed scenes to a specified final outcome. The task difficulty is controlled by varying the number of missing events, ranging from one to three:

- **1-Hop:** The model predicts a single future event that directly links the observed scenes to the final one. This corresponds to a standard NEP.
- **2-Hop:** The model infers a sequence of two consecutive future events, requiring a short chain reasoning process that sequentially connects the observed scenes to the final event.
- **3-Hop:** The model predicts three consecutive future events, significantly increasing task complexity by necessitating deeper causal reasoning across a longer temporal span.

**Future Event Prediction—Interpolation.** The interpolation subtask introduces a complementary challenge wherein the model must infer multiple non-consecutive future events, given a set of partially observed scenes that include intermediate anchor events. Rather than constructing a continuous sequence – as in extrapolation – this task demands the model interpolate across disjoint glimpses of future events. It emphasizes reasoning over causal continuity and temporal coherence amid fragmentary observation, as illustrated in Figure 4.

#### 4.2 DETAILS OF QUESTION-ANSWER PAIR GENERATION

Designing high-quality questions and answer choices for FutureBench presents a non-trivial challenge, as it demands capturing the nuanced temporal logic embedded in each narrative. To scale the generation of QA pairs, we adopt a LLM-based generation pipeline. We employ GPT-4 (text-only mode) to generate QA pairs from detailed video annotations. Each video is accompanied by rich textual metadata, including a synopsis, segment-level scene descriptions, a specification of the observed scenes (i.e., the initial context), and a description of the final scene (i.e., the target outcome). We then prompt GPT-4 using a structured template designed to emulate a human question-setter. The prompt instructs GPT-4 to formulate a question that probes for the missing future events and to generate a correct answer along with several plausible yet incorrect distractors. **To ensure that the question requires genuine reasoning**, the prompt explicitly references the need to achieve a final outcome and is carefully crafted to prevent shortcut solutions – for examples, by avoiding lexical overlap between the correct answer and question, or easily dismissible distractors. Additionally, the distractor choices are constructed to be commonsense-plausible within the thematic context of the video but logically inconsistent with the outcome trajectory, thereby increasing task difficulty. An illustrative example of this process is shown in Figure 4, and the full prompt used for GPT-4 is provided in Appendix E. As a result, FutureBench comprises a total 1056 carefully curated QA pairs spanning both extrapolation and interpolation subtasks. To assess the benchmark’s quality and highlight both visual perception and temporal reasoning, we evaluate a strong reasoning model, o4-mini, on the text-only version of questions, excluding any visual input. The model achieves an accuracy of 32.0%, suggesting that **even advanced reasoning capabilities alone are insufficient for consistently solving the tasks**. This finding reinforces the critical role of visual perception in solving future event prediction in FutureBench. More details regarding dataset distribution and the discussion of human-in-the-loop quality review in B.3.

Table 2: **Performance comparison of different instruction tuning strategies.** G-Avg. and T-Avg. represent the average performances of all general and temporal benchmarks, respectively. Instruct represents the original performances without additional training.

Models	General Benchmark				Temporal Benchmark				
	VMME <sub>(w/o sub)</sub>	MVB	LVB <sub>val</sub>	G-Avg.	TB	TC	SB-R1	FB	T-Avg.
<b>Qwen2.5-VL-3B-Instruct</b>									
Instruct	55.7	63.8	52.2	57.2	30.8	<b>69.3</b>	33.2	49.9	45.8
SFT	55.8	62.8	50.4	56.3	34.3	61.5	<b>35.7</b>	<b>61.1</b>	48.2
CFT	55.6	63.1	50.9	56.5	32.6	68.5	34.6	50.1	46.5
Distill	56.2	64.5	<b>53.5</b>	<b>58.1</b>	33.9	69.1	33.6	57.2	48.4
Mix	<b>56.6</b>	<b>64.6</b>	52.4	57.9	<b>34.8</b>	66.5	<b>35.7</b>	56.9	<b>48.5</b>
<b>Qwen2.5-VL-7B-Instruct</b>									
Instruct	59.8	65.3	55.9	60.3	35.4	73.8	37.1	52.6	49.7
SFT	59.2	66.5	53.4	59.7	<b>39.9</b>	69.9	39.1	61.3	52.6
CFT	58.9	65.3	54.2	59.5	35.2	74.1	<b>39.8</b>	55.8	51.2
Distill	<b>60.6</b>	<b>66.7</b>	<b>56.3</b>	<b>61.2</b>	35.9	<b>75.1</b>	37.0	59.5	51.9
Mix	59.6	66.4	53.7	59.9	38.2	72.9	38.5	<b>63.4</b>	<b>53.3</b>

## 5 EXPERIMENT

### 5.1 COMPARISON WITH OTHER VIDEO INSTRUCTION TUNING TASKS

To investigate the effectiveness of NEP as a learning task, we fine-tune Qwen2.5-VL-7B-Instruct on NEP and compare its performance against models trained on three prior instruction tuning tasks: captioning, multi-choice question answering (**MCQA**), and open-ended question answering (**OEQA**). For fairness, all models are trained on a dataset of equal size using 3K samples. For the captioning, MCQA and OEQQA, we use the data constructed by LLaVA-Video-178K (Zhang et al., 2023).

To comprehensively evaluate model performance, we consider two groups of benchmarks. First, we assess general video understanding on three widely-used benchmarks that are not specifically designed to test temporal reasoning: VideoMME<sub>(w/o sub)</sub> (**VMME**) (Fu et al., 2024), MVBench (**MVB**) (Li et al., 2024d), and LongVideoBench<sub>val</sub> (**LVB**) (Wu et al., 2024). Second, to examine temporal understanding and reasoning capabilities, we evaluate on four temporally-focused benchmarks: TemporalBench (**TB**) (Cai et al., 2024), TempCompass (**TC**) (Liu et al., 2024b), SeedBench-R1 (**SB-R1**) (Chen et al., 2025), and our proposed FutureBench (**FB**). These benchmarks challenge models to make complex temporal understanding and reasoning. For all evaluations, we use 32 frames from the video as the input by default. Detailed training and evaluation descriptions can be found in Appendix B.2.

**Next-event prediction enhances temporal reasoning without sacrificing general video understanding.** As shown in Table 1, models trained on the NEP task with partially observed video demonstrate substantial improvements on temporal benchmarks compared to those trained on Captioning, MCQA, and OEQQA tasks with the full observed video. Notably, NEP-trained models also maintain competitive performance on general benchmarks, underscoring the superiority and compatibility of the NEP task. These findings suggest that NEP not only strengthens a model’s ability to reason over temporal sequences but does so without compromising its overall comprehension abilities. NEP serves as an effective learning signal that promotes both visual perception and temporal reasoning with minimal trade-offs in general performance.

**Deductive reasoning via next-event prediction yields greater improvements on temporal benchmarks compared to inductive (video Q&A) and abductive (previous-event prediction) reasoning.** Video Q&A, next-event prediction, and previous-event prediction can be deemed three classical forms of logical reasoning—induction, deduction, and abduction, respectively (Douven, 2021; Cheng et al., 2024). An illustration is provided in Figure 17 in the appendix. To study the relative efficacy of these reasoning types, we fine-tune the Qwen2.5-VL-7B-Instruct model using the same training set of 3K samples, modifying only the task formulation to align with each reasoning. The results presented in Table 5 indicate that the deduction task, next event prediction, yields significantly greater performance on temporal benchmarks compared to induction and abduction tasks. In contrast to induction and abduction, deduction often involves the deliberate application of abstract logical principles. Such reasoning tends to be more cognitively demanding and typically necessitates targeted learning and structured practice (Goswami, 2010; Behfar & Okhuysen, 2018).

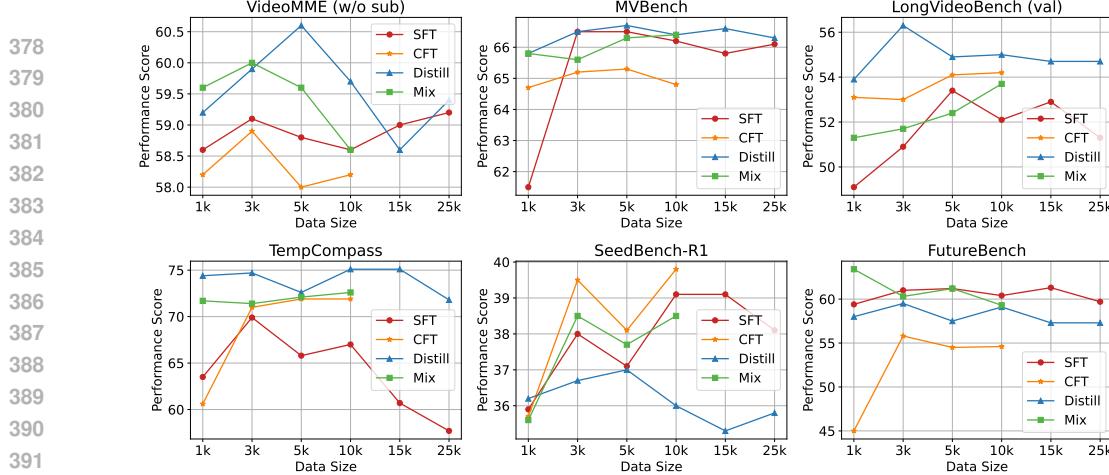


Figure 5: **Performance comparison of different data scales for SFT, CFT, Distill, and Mix tuning on Qwen2.5-VL-7B-Instruct.** The top showcases the curves for general benchmarks, and the bottom showcases the curves for temporal benchmarks.

## 5.2 COMPARISON OF INSTRUCTION TUNING STRATEGIES

To further explore effective strategies for training on the NEP task, we compare four instruction tuning approaches introduced in Section 3.3: supervised fine-tuning (SFT), critique fine-tuning (CFT), distillation (Distill), and mix tuning (Mix). We conduct experiments on both Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct, evaluating each strategy across general and temporal video benchmarks. Additionally, we study the impact of training set size by scaling SFT and Distill from 1K to 25K samples, and CFT and Mix from 1K to 10K samples. For the Mix setting, we fix a total tuning budget and then combine examples from different supervision types in equal proportions (e.g., an equal number of SFT-, Distill-, and CFT-style NEP instances under the same total number of samples). During training, mini-batches are sampled from this unified mixed pool.

**SFT serves as a simple but effective strategy for NEP training.** As shown in Table 2, simple SFT yields substantial gains on temporal benchmarks, demonstrating its efficacy for NEP. While CFT and Distill also contribute notable improvements, they rely on additional annotations or feedback from auxiliary LLMs, making them less efficient in comparison to SFT. Importantly, Mix strategy achieves the highest average performance on temporal benchmarks, effectively combining the strengths of all tuning methods. We hypothesize that this is due to the complementary nature of supervision signals: SFT provides direct supervision via ground-truth next events, while CFT and Distill introduce richer semantic feedback through model-generated guidance. This diversity likely enables the model to better generalize in temporal prediction tasks.

**Effect of training data scale.** As illustrated in Figure 5, increasing the NEP training data size beyond 5K samples does not uniformly improve performance across tuning strategies and, in some cases, even leads to slight degradation on both general and temporal benchmarks. We think this saturation is due to several factors. *First*, NEP is specifically designed to enhance temporal reasoning rather than to inject additional general world knowledge. Therefore, we do not expect NEP-only training to consistently improve general-purpose benchmarks as the data scale grows; in this regime, maintaining comparable general performance (i.e., avoiding degradation) is already a desirable outcome. *Second*, to cleanly isolate the effect of the training task itself, we construct NEP data from the intersection of video corpora that have already been used for other training tasks, including captioning, open-ended QA (OEQA), and multiple-choice QA (MCQA). Using this core set of videos allows a controlled, task-level comparison under a fixed video corpus, but, in turn, it also means that NEP is trained on a restricted set of video sources. This reduced diversity may limit the room for scaling: when additional NEP samples come from essentially the same narrow distribution,

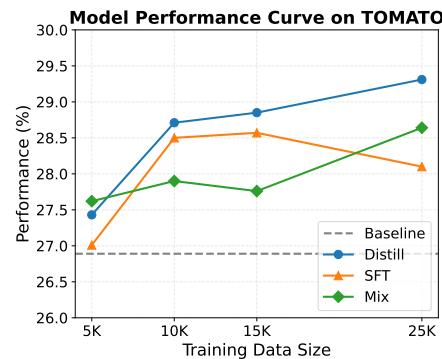


Figure 6: **Data-scaling behavior on the TOMATO temporal reasoning benchmark.**

Table 3: **Ablation on data construction strategies for NEP tuning.** G-Avg. and T-Avg. represent the average performances of general and temporal benchmarks.

	General Benchmark				Temporal Benchmark				
	VMME <sub>(w/o sub)</sub>	MVB	LV <sub>B</sub> <sub>val</sub>	G-Avg.	TB	TC	SB-R1	FB	T-Avg.
Qwen2.5-VL-7B-Instruct	59.8	65.3	<b>55.9</b>	<b>60.3</b>	35.4	<b>73.8</b>	37.1	52.6	49.7
SFT w/o Caption Analysis	<b>60.9</b>	65.6	52.9	59.8	37.1	69.4	37.6	59.5	50.9
SFT w. Caption Analysis	59.2	<b>66.5</b>	53.4	59.7	<b>39.9</b>	69.9	<b>39.1</b>	<b>61.3</b>	<b>52.6</b>

Table 4: Comparison of different training tasks on TOMATO across reasoning types.

Model	Count	Direction	Rotation	Shape&Trend	Velocity&Freq	Visual Cues	Overall
NEP-Distill	32.53	28.54	24.48	33.63	26.67	34.29	29.31 ( <b>+2.42</b> )
NEP-CFT	36.99	31.76	19.58	31.39	19.52	40.00	29.04 ( <b>+2.15</b> )
NEP-MIX	31.51	29.53	26.92	30.49	23.33	28.57	28.64 ( <b>+1.75</b> )
NEP-SFT	37.50	30.52	20.80	27.58	21.42	36.43	28.57 ( <b>+1.68</b> )
OEQA	41.10	26.05	18.53	31.39	21.43	42.86	28.50 ( <b>+1.61</b> )
MCQA	39.73	27.30	19.93	26.46	21.90	31.43	27.63 ( <b>+0.74</b> )
Qwen2.5-VL-7B	34.93	28.29	21.68	23.32	18.10	44.29	26.89
Captioning	34.59	26.80	22.73	25.56	17.14	44.29	26.82 ( <b>-0.07</b> )

increasing the data size brings smaller gains. *Third*, our experimental protocol deliberately trains NEP in isolation and compares it against models trained solely on MCQA or OMQA, so that we can study “training task vs. training task” under a fixed budget. Because most evaluation benchmarks are formulated in QA format, this design places NEP at a slight disadvantage: NEP-only tuning introduces a shift between the NEP objective and the downstream QA-style usage. For the second and third points, in a practical recipe for pushing state-of-the-art performance, one would collect as much diverse data as possible, combine NEP with QA and captioning data, and carefully schedule when to use which supervision, rather than training with NEP alone. *Finally*, existing temporal benchmarks may under-represent the full benefits of NEP. As discussed in the recent temporal reasoning multi-modal evaluation benchmark TOMATO (Shangguan et al., 2024), “existing widely-used temporal reasoning video benchmarks (Liu et al., 2024b; Chen et al., 2024a; Li et al., 2024e) allow models to exploit shortcuts, enabling them to answer correctly using a single, few, or out-of-order frames”. In contrast, Figure 6 shows that, on TOMATO, NEP exhibits a more positive data-scaling trend: more NEP training data generally yields better temporal performance, although with some non-monotonic fluctuations.

**Effect of Caption Analysis.** The caption analysis step is designed to improve both the quality and the challenge level of constructed samples. Without this step, we fall back to selecting a sentence boundary from the caption as the text split point. As shown in Table 3, applying SFT to data constructed in a simple way without caption analysis already brings a modest improvement on temporal benchmarks over the untuned baseline (T-Avg.: 49.7 → 50.9), while largely preserving performance on general benchmarks. Furthermore, applying caption analysis on top of the same video corpus yields consistently larger gains in temporal reasoning, increasing T-Avg. from 50.9 to 52.6. These results demonstrate that caption analysis is an effective component of the NEP data construction pipeline.

### 5.3 ASSESSING FINE-GRAINED VISUAL TEMPORAL REASONING CAPABILITIES

Beyond the four temporal benchmarks (TempCompass, TemporalBench, SeedBench-R1, and our FutureBench) in Section 5.1, we further ask whether *next-event prediction* (NEP) improves *genuinely* temporal reasoning rather than exploiting static shortcuts. To this end, we evaluate our models on TOMATO (Shangguan et al., 2024), a recent benchmark explicitly designed to test fine-grained visual temporal reasoning in multimodal foundation models. The benchmark contains 1,484 human-curated multiple-choice questions built on 1,417 videos, including 805 self-recorded or synthesized clips. It covers six temporally focused tasks: *Rotation* (global rotation direction), *Direction* (motion direction), *Velocity & Frequency* (changes in speed or repetition rate), *Shape & Trend* (trajectory shape or overall trend), *Visual Cues* (subtle signals for ordering or timing), and *Action Count* (number of repetitions).

We follow the official TOMATO evaluation pipeline. For each example, the script samples a fixed number of frames, constructs a multiple-choice prompt with the question and answer options, and queries the model once. We then use the official parser to extract the predicted option and compute

486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

**Table 5: Performance comparison of inductive, deductive, and abductive tasks on temporal benchmarks.** PEP: Previous Event Prediction.

	Temporal Benchmark			
	TB	TC	SB-R1	FB
Inductive (Video QA)	36.6	74.0	35.4	58.8
Deductive (NEP)	<b>38.6</b>	<b>74.7</b>	<b>39.5</b>	<b>61.3</b>
Abductive (PEP)	38.0	66.2	31.2	55.1

**Table 6: Performance comparison of SFT and GRPO with NEP.** G-Avg.: average performance of general benchmarks. Interp.: Interpolation task.

	General G-Avg.	FutureBench			
		1-Hop	2-Hop	3-Hop	Interp.
Instruct	<b>60.3</b>	56.1	57.5	49.8	50.5
NEP w. SFT	59.7	67.6	64.2	57.7	59.3
NEP w. GRPO	58.2	<b>83.8</b>	<b>81.3</b>	<b>62.7</b>	<b>65.2</b>

accuracy, and report scores aggregated by *reasoning type* (the six temporal tasks above). This decomposition allows us to examine whether NEP preferentially improves particular kinds of motion reasoning. For each task, we report the **best** accuracy achieved over training-set sizes from 5K to 25K samples (same as in the previous experiments), as summarized in Table 4. Among all configurations, NEP with the Distill training strategy yields the largest improvement, suggesting it most effectively strengthens temporal reasoning. Consistent with Table 1, training on the multiple-choice QA task leads to lower absolute scores than open-ended video QA on the TOMATO benchmark.

#### 5.4 REINFORCEMENT LEARNING WITH NEXT-EVENT PREDICTION

Reinforcement learning (RL) represents an alternative and essential learning paradigm for enhancing reasoning capabilities. To systematically examine the impact of RL-based training of NEP on both general and temporal video understanding, we construct a dedicated training set comprising 2,000 multi-choice QA pairs. This training set is generated using the same pipeline as FutureBench, but is derived from the V1-33K video dataset and restricted to 1-hop and 2-hop extrapolation tasks. Consequently, the 3-hop extrapolation task is treated as an out-of-distribution (OOD) setting, designed to assess model generalization to longer, unseen causal chains. Similarly, the interpolation task (Interp.) presents an additional OOD challenge, requiring the model to reason over fragmented future context. In this experiment, we train the Qwen-2.5-VL-7B-Instruct using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with the outcome supervision and evaluate its performance across both general and temporally-focused video benchmarks.

**RL generalizes well on FutureBench but degrades performance on general benchmarks.** As shown in Table 6, the GRPO-trained model demonstrates strong performance improvement on in-distribution tasks and generalizes well to OOD tasks, including 3-hop questions and interpolation tasks. These results underscore the effectiveness of RL training in the future event prediction task. However, it is also notable that the RL-trained model suffers from non-trivial performance degradation on general video understanding benchmarks. This suggests that while RL training promotes a reasoning style suited for future event prediction, it may pose inductive biases that hinder generalizability to tasks not requiring future-oriented prediction. Furthermore, we observe instances of reward hacking, wherein RL training with multi-choice QA and outcome supervision may encourage models to exploit superficial patterns, such as lexical similarity between answer options and the question text, to arrive at correct answers. Such behavior deviates from our initial motivation and this shortcut undermines the intended objective of next-event prediction, which is to foster integrated visual perception and causal reasoning. Given these limitations, *we highlight that SFT-style training remains a simple yet efficient approach for training on NEP.*

## 6 CONCLUSION

In this work, we propose next-event prediction, a self-supervised learning task designed specifically to improve temporal reasoning capabilities in MLLMs. By dividing videos into past and future frames, NEP forces models to predict unseen future events, enabling models to implicitly build robust internal representations of causal and narrative dynamics. To study NEP and facilitate research in this area, we created V1-33K, a large dataset of approximately 33,000 video instances that cover a wide range of real-world scenarios and temporal complexities. Furthermore, we proposed FutureBench, a comprehensive benchmark that assesses models’ ability to generate logically coherent and causally consistent future event predictions. Experiments show that incorporating NEP significantly improves MLLMs’ temporal reasoning capabilities while maintaining their performance on traditional video understanding tasks. We believe that NEP lays a foundation for advancing temporal understanding in MLLMs, bridging the gap between static visual description and temporal event inference.

540 REFERENCES  
541

542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
543 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.  
544 *arXiv preprint arXiv:2303.08774*, 2023.

545 Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar  
546 Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models  
547 enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

548 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
549 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

550 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,  
551 Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan,  
552 Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng,  
553 Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv  
554 preprint arXiv:2502.13923*, 2025.

555 Apoorva Beedu, Harish Haresamudram, Karan Samel, and Irfan Essa. On the efficacy of text-based  
556 input modalities for action anticipation. *arXiv preprint arXiv:2401.12972*, 2024.

557 Kristin Behfar and Gerardo A Okhuysen. Perspective—discovery within validation logic: Deliberately  
558 surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry.  
559 *Organization Science*, 29(2):323–340, 2018.

560 Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu  
561 Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding  
562 for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

563 Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank  
564 Wang. Rextime: A benchmark suite for reasoning-across-time in videos, 2024a. URL <https://arxiv.org/abs/2406.19392>.

565 Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Lu Qiu, Ying Shan, and Xihui Liu. Exploring the effect  
566 of reinforcement learning on video understanding: Insights from seed-bench-r1. *arXiv preprint  
567 arXiv:2503.24376*, 2025.

568 Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong  
569 Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning  
570 for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer  
571 Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

572 Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang  
573 Li, Zheng Li, Yifan Gao, Xian Li, et al. Inductive or deductive? rethinking the fundamental  
574 reasoning abilities of llms. *arXiv preprint arXiv:2408.00114*, 2024.

575 Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost  
576 in time: A new temporal benchmark for videollms, 2025. URL <https://arxiv.org/abs/2410.07752>.

577 Igor Douven. Abduction. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.  
578 Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

579 Chaoyou Fu, Yuhai Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu  
580 Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation  
581 benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

582 Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future:  
583 A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International  
584 Conference on Computer Vision*, pp. 5562–5571, 2019.

585 Usha Goswami. Inductive and deductive reasoning. *The Wiley-Blackwell handbook of childhood  
586 cognitive development*, pp. 399–419, 2010.

594 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,  
 595 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms  
 596 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

597

598 Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers.  
 599 *arXiv preprint arXiv:2212.10071*, 2022.

600

601 Mehrdad Hosseinzadeh and Yang Wang. Video captioning of future frames. In *Proceedings of the*  
 602 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 980–989, January  
 603 2021.

604

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

608

609 Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action  
 610 prediction. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland,  
 September 6–12, 2014, Proceedings, Part III 13*, pp. 689–704. Springer, 2014.

611

612 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan  
 613 Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint*  
 614 *arXiv:2408.03326*, 2024a.

615

616 Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin  
 617 Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model.  
 618 *arXiv preprint arXiv:2410.05993*, 2024b.

619

620 Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zeyun Ma, and Chunyuan Li.  
 621 Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv*  
 622 *preprint arXiv:2407.07895*, 2024c.

623

624 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,  
 625 Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In  
 626 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.  
 627 22195–22206, 2024d.

628

629 Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs:  
 630 A diagnostic dataset for temporal concept understanding of video-language models. In *European*  
 631 *Conference on Computer Vision*, pp. 331–348. Springer, 2024e.

632

633 Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey  
 634 of multimodel large language models. In *Proceedings of the 3rd International Conference on*  
 635 *Computer, Artificial Intelligence and Control Engineering*, pp. 405–409, 2024.

636

637 Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning  
 638 united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*,  
 639 2023.

640

641 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*  
 642 *neural information processing systems*, 36:34892–34916, 2023.

643

644 Xin Liu, Chao Hao, Zitong Yu, Huanjing Yue, and Jingyu Yang. From recognition to prediction:  
 645 Leveraging sequence reasoning for action anticipation. *ACM Transactions on Multimedia*  
 646 *Computing, Communications and Applications*, 20(11):1–19, 2024a.

647

648 Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. When the future becomes  
 649 the past: Taming temporal correspondence for self-supervised video representation learning. In  
 650 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24033–24044, 2025.

651

652 Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and  
 653 Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*,  
 654 2024b.

648 Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt:  
 649 Towards detailed video understanding via large vision and language models. *arXiv preprint*  
 650 *arXiv:2306.05424*, 2023.

651 Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging  
 652 the present to anticipate the future in videos. In *Proceedings of the IEEE/CVF Conference on*  
 653 *Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

654 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 655 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow  
 656 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–  
 657 27744, 2022.

658 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
 659 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
 660 models from natural language supervision. In *International conference on machine learning*, pp.  
 661 8748–8763. PMLR, 2021.

662 MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit  
 663 Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv*  
 664 *preprint arXiv:1412.6604*, 2014.

665 Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman  
 666 Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models.  
 667 *arXiv preprint arXiv:2410.23266*, 2024.

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

671 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,  
 672 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlfh framework. *arXiv preprint*  
 673 *arXiv: 2409.19256*, 2024.

674 Alexandros Stergiou and Dima Damen. The wisdom of crowds: Temporal progressive attention for  
 675 early action prediction. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2023.

676 Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An,  
 677 Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey.  
 678 *arXiv preprint arXiv:2312.17432*, 2023.

679 Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL <https://arxiv.org/abs/2312.11805>.

680 Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin  
 681 Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang,  
 682 Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding,  
 683 Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng  
 684 Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie,  
 685 Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan  
 686 Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan  
 687 Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu,  
 688 Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles,  
 689 Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yeqie Wang, Yibo Liu, Yibo Miao, Yidao  
 690 Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin  
 691 Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang,  
 692 Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025. URL  
 693 <https://arxiv.org/abs/2504.07491>.

694 Vinh Tran, Yang Wang, Zekun Zhang, and Minh Hoai. Knowledge distillation for human action  
 695 anticipation. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2518–2522.  
 696 IEEE, 2021.

702 Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics.  
 703 *Advances in neural information processing systems*, 29, 2016.  
 704

705 Yubo Wang, Xiang Yue, and Wenhui Chen. Critique fine-tuning: Learning to critique is more effective  
 706 than learning to imitate, 2025. URL <https://arxiv.org/abs/2501.17703>.  
 707

708 Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context  
 709 interleaved video-language understanding. *Advances in Neural Information Processing Systems*,  
 710 37:28828–28857, 2024.  
 711

712 Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-  
 713 answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on  
 Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.  
 714

715 Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language  
 716 model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.  
 717

718 Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu,  
 719 Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on  
 720 the evaluation of large multimodal models, 2024a. URL <https://arxiv.org/abs/2407.12772>.  
 721

722 Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video  
 723 instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.  
 724

725 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyuan Luo, Zhangchi Feng, and  
 726 Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Pro-  
 ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3:  
 System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics.  
 727 URL <http://arxiv.org/abs/2403.13372>.  
 728

729 Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong.  
 730 Easyr1: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyR1>, 2025.  
 731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 A APPENDIX: RELATED WORK  
757758 A.1 VIDEO INSTRUCTION-TUNING OF MLLMs  
759

760 The fusion of vision and language in large models has advanced rapidly from image-focused models  
761 like CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023) to recent video-language models that  
762 interpret dynamic visual content leveraging the advanced ability of LLMs (Liang et al., 2024; Tang  
763 et al., 2023). Early approaches adapted image-based techniques by fine-tuning LLMs with an extended  
764 visual encoder on video frames for observational tasks, such as captioning and question answering;  
765 this process is also known as video instruction tuning. For instance, Video-ChatGPT (Maaz et al.,  
766 2023) was trained on a large custom 100k video-instruction (Q&A) dataset derived from sources  
767 like ActivityNet-200 to enable detailed video conversation and reasoning. Video-LLaVA (Lin et al.,  
768 2023) used mixed image/video datasets focusing on unified visual QA across modalities. Video-  
769 LLaMA (Zhang et al., 2023) utilized large-scale video/image caption datasets (e.g., Webvid-2.5M,  
770 LLaVA-CC3M) for pre-training and various instruction datasets (MiniGPT-4, LLaVA, VideoChat)  
771 for fine-tuning, targeting video-to-text generation and audio-visual instruction following. LLaVA-  
772 OneVision (Li et al., 2024a) leveraged large-scale image instruction data (covering QA, OCR, Math,  
773 etc.) aiming for cross-scenario transfer to video understanding tasks like referring analysis, and  
774 LLaVA-NeXT-Interleave (Li et al., 2024c) compiled the diverse M4-Instruct dataset (1.18M samples  
775 across 41 datasets) for interleaved multi-modal QA and captioning across images, video frames, and  
776 3D views. Even models built by companies, like Qwen2.5-VL (Bai et al., 2025), Kimi-VL (Team  
777 et al., 2025) and Aria (Li et al., 2024b) primarily build upon datasets focused on description, QA, or  
778 dialogue generation based on observed content.

779 However, their training data and objectives are predominantly observational, describing or explaining  
780 visible content, rather than predictive. Our work differs by introducing a predictive objective, next  
781 event prediction, to explicitly train the model’s temporal reasoning abilities. This aligns with the goal  
782 of modeling world dynamics, extending beyond static understanding of frames to reasoning about  
783 how scenes evolve over time. While prior instruction-tuned MLLMs were not primarily designed for  
784 complex temporal reasoning, their advanced video perception abilities provide a strong foundation,  
785 paving the way for our work to further explore and advance temporal reasoning mechanisms in  
786 MLLMs.

787 A.2 FUTURE PREDICTION IN COMPUTER VISION  
788

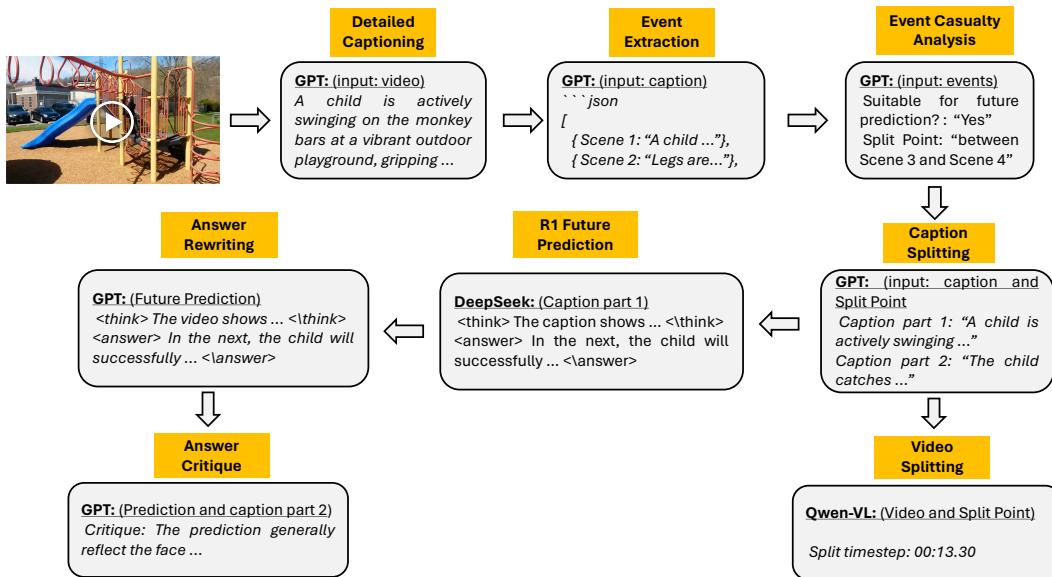
789 Anticipating the future has been studied in computer vision in various forms. A significant body of  
790 work focuses on action prediction tasks, which frame future prediction as a classification problem.  
791 The goal is to improve video encoder by predicting the next discrete action label, typically from a  
792 predefined set of (verb, noun) pairs, based on observed video frames (Miech et al., 2019). Within this  
793 paradigm, various methodologies have been explored to improve predictive accuracy. For instance,  
794 Liu et al. (2024a) define the task into action recognition and introduce a Next Action Prediction  
795 (NAP) objective, analogous to next-token prediction in language models, to explicitly model the  
796 statistical correlations between consecutive actions. Other approaches focus on incorporating richer  
797 input modalities. Beedu et al. (2024) propose the M-CAT model, which demonstrates that using  
798 text-based descriptions of objects and actions as an additional input modality can provide crucial  
799 context, thereby enhancing the model’s ability to classify the correct future action. Methodological  
800 improvements have also been achieved through advanced training strategies. For example, Tran et al.  
801 (2021) utilize a knowledge distillation framework where a student model, which only observes past  
802 frames, is trained to predict the future-aware representations of a teacher model that has access to  
803 both past and future frames.

804 Another major paradigm uses future prediction as a self-supervised objective for representation  
805 learning. Here, the goal is not the prediction itself, but to pre-train a powerful, general-purpose video  
806 encoder for various downstream applications. For example, Liu et al. (2025) propose a framework  
807 that reconstructs masked patches of the current frame in a latent space. To guide this reconstruction,  
808 it employs a “sandwich sampling” strategy that leverages temporal correspondence from both past  
809 and future frames, explicitly avoiding the low-level detail required by pixel-space reconstruction (Liu  
et al., 2025). Similarly, Assran et al. (2025) introduce V-JEPA 2, a joint-embedding predictive  
architecture that learns a world model by predicting the representations of masked video segments in

810 a learned feature space. While V-JEPA 2 achieves state-of-the-art results on the downstream task of  
 811 action anticipation, its primary contribution is the learned powerful video encoder.  
 812

813 Distinct from both classification and representation learning, Hosseinzadeh & Wang (2021) propose  
 814 a method for video captioning of future frames where the goal is to generate a sentence describing  
 815 a likely future event. They propose a disjoint approach: in first stage, it first predicts the video  
 816 representation of unobserved frames; and in the second stage, it feeds these predicted features into a  
 817 dedicated captioning translator to convert representation into the final textual description.  
 818

819 These works typically define the “future” at short horizons, e.g., the next frame or next action, and  
 820 optimize low-level objectives, e.g., representation learning, which has proven effective for video  
 821 encoder pretraining. Our work is distinct in targeting high-level, semantic future-event prediction  
 822 expressed in natural language. This NEP task requires MLLMs to perceive temporally grounded  
 823 visual evidence from the vision module, integrate it with pretrained commonsense knowledge in  
 824 the LLM module, and generate open-ended outcomes in natural language. Because the goal differs,  
 825 so does the focus of optimization: prior methods mainly update the visual encoder, we freeze the  
 826 encoder and fine-tune the vision–language projector and the LLM, enhancing multimodal temporal  
 827 reasoning.  
 828



## B DETAILED DATA CONSTRUCTION PIPELINE

### B.1 V1-33K CONSTRUCTION

849 **Fact Translation.** In this initial stage, visual content is translated into a textual format to serve as  
 850 the foundation for further processing. For every video, we use a Vision-Language Model (VLM) to  
 851 generate a detailed caption that comprehensively describes the visual facts. This conversion from  
 852 visual to textual data ensures that the strong text-based reasoning capabilities of open-source large  
 853 language models (LLMs) can be leveraged.  
 854

855 **Analysis.** Given the fact that current models exhibits stronger reasoning capabilities when working  
 856 with text, we feed the detailed captions into a LLM. The LLM performs two critical tasks:  
 857

- *Scene Identification:* It dissects the caption to extract and delineate distinct scenes.
- *Causal Analysis:* It evaluates the causal relationships between scenes and identifies an optimal  
 858 split point where the context from preceding events is strong enough to predict what comes next.  
 859

860 This step establishes a structured understanding of the video, which is crucial for effective segmenta-  
 861 tion. The prompt used in this step can be found in Appendix E.2.  
 862

864     **Segmentation.** Using the optimal split point determined during the Analysis stage, we partition  
 865     both the original video and its caption into two parts. The first part of the video, which contains the  
 866     initial events, serves as a clear input for the video reasoning model, ensuring that the video reasoning  
 867     is based on established facts. The second segment is reserved as the ground truth for evaluating the  
 868     model’s predictions. **The prompt used in this step can be found in Appendix E.3.**  
 869

870     **Reasoning & Critique.** One promising approach to rapidly enhance video reasoning is through  
 871     Long CoT supervised fine-tuning. In our dataset, we leverage the output of a text reasoning model to  
 872     facilitate this process. Specifically, the text reasoning model (DeepSeek-R1) processes the first part of  
 873     the caption, recording its reasoning process and generating predictions for future events. Recognizing  
 874     that textual reasoning can sometimes introduce errors, we subsequently employ an additional LLM  
 875     to critically evaluate both the reasoning process and the resulting predictions. This approach draws  
 876     inspiration from recent advances in critique fine-tuning (CFT), where models learn to critique noisy  
 877     responses, specifically the reasoning and predictions, rather than simply imitating them through SFT.  
 878     By doing so, we ensure that only robust reasoning informs the final training of the MLLM, ultimately  
 879     boosting its overall performance. **The prompts used in this step can be found in Appendix E.4 and**  
 880     **E.5.**

881     The data processing pipeline is outlined below. We employ DeepSeek-R1 [Guo et al. \(2025\)](#) for  
 882     the Future Prediction step and Qwen2.5-VL-72B-Instruct for Video Splitting, while using the O3-  
 883     mini [Achiam et al. \(2023\)](#) for all other steps. The prompts used at each stage are critical for  
 884     high-quality data processing. We have made efforts in manually testing a wide range of hand-written  
 885     prompts and playing with the API.

886     Table 7: Statistics and distribution of data source for Extrapolation and Interpolation in FutureBench.  
 887     #Total indicates the total size of each subset.

Data Source	Extrapolation			Interpolation
	1-Hop	2-Hop	3-Hop	
#Total	173	193	201	489
YouTube	48.0%	37.3%	45.3%	51.9%
ActivityNet	23.1%	31.6%	24.9%	23.5%
YouCook2	11.6%	10.4%	10.0%	8.2%
NextQA	8.7%	10.4%	10.0%	8.2%
Charades	8.6%	10.3%	9.8%	8.2%

## B.2 FUTUREBENCH DETAILS

901     We discuss the details of FutureBench construction in Section 4.2. Note that the videos used in  
 902     FutureBench have no overlap with V1-33K to ensure fair evaluation despite the same curation pipeline.  
 903     FutureBench also involves videos from diverse sources. The final statistics of FutureBench and  
 904     distribution of the data source are shown in Table 7.

## B.3 HUMAN-IN-THE-LOOP QUALITY REVIEW

908     Following automatic generation, all QA items undergo a verification and filtering process. Samples  
 909     deemed too trivial – **such as those with answers directly inferable from a single frame, or with**  
 910     **implausible distractors using simple commonsense** – are discarded. QA pairs requiring minor  
 911     corrections are edited to ensure semantic coherence and alignment with the underlying video narrative.  
 912     This human-in-the-loop review process allows us to maintain high annotation quality while leveraging  
 913     GPT-4 to scale data generation efficiently.

914     As a result, FutureBench comprises a total 1056 carefully curated QA pairs spanning both extrapolation  
 915     and interpolation subtasks. To assess the benchmark’s quality and highlight both visual  
 916     perception and temporal reasoning, we evaluate a strong reasoning model, o4-mini, on the text-only  
 917     version of questions, excluding any visual input. The model achieves an accuracy of 32.0%, suggesting  
 918     that **even advanced reasoning capabilities alone are insufficient for consistently solving the**

918 **tasks.** This finding reinforces the critical role of visual perception in solving future event prediction  
 919 in FutureBench  
 920

## 921 C QUALITATIVE EVALUATION OF DATA AND BENCHMARK

### 922 C.1 QUALITATIVE STUDY OF FUTUREBENCH

923 In here we provide the qualitative studies of the FutureBench.

924 For the sample shown in Figure 8, option A is correct because after the observed scenes (where the  
 925 digits and straight lines are drawn), the next logical step in the timeline is to modify the drawn ‘2’ s  
 926 by adding a ‘C’ shape on top of each one. This action initiates the transformation of the numbers into  
 927 elements that will later become detailed bird-like figures, fitting neatly between the earlier steps and  
 928 the final completed image.

929 For the sample shown in Figure 9, option A correctly predicts the intermediate events: a chaotic  
 930 moment where the man tumbles on the countertop, followed by the woman’s concerned reaction and  
 931 her cracking an egg into a bowl, which bridges logically between the observed scenes and the final  
 932 scene of presenting the cake. Option B distracts by describing a calm, uneventful sequence that does  
 933 not match the dynamic and slightly chaotic tone of the video. Option C introduces conflict that is not  
 934 supported by the context of collaborative, playful baking, and Option D describes unrelated actions  
 935 (ignoring the kitchen and chopping vegetables) that do not bridge to the final cake presentation.

936 For the sample shown in Figure 10, option D correctly bridges the gap by introducing the sauteing  
 937 of fresh ingredients, which establishes the cooking of the dish, followed by the careful plating that  
 938 logically leads into the final garnished presentation. Option A incorrectly emphasizes rolling and  
 939 cutting dough again, repeating events already observed and omitting the important cooking phase.  
 940 Option B ignores the gradual progression of cooking and assembling the dish by suggesting an  
 941 immediate processing of the dough. Option C, although mentioning dough cutting and feeding, lacks  
 942 the essential steps of sauteing and final plating that would lead to the garnished final scene.

943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

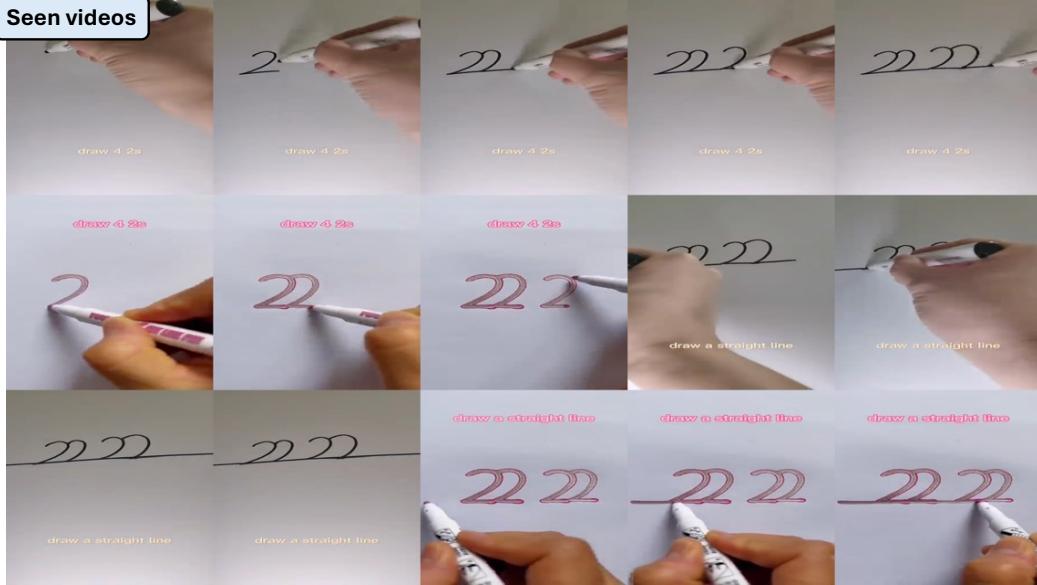
1000

1001

1002

1003

1004

**Seen videos**

A: A hand holding a black marker draws a 'C' shape on top of each '2' to begin transforming the numbers into bird-like figures.

B: A hand holding a pink marker draws two 'W' shapes beneath the figures to indicate the start of wing formation.

C: A hand holding a black marker draws two curved lines on each '2', forming bird-like shapes with implied bodies.

D: A hand holding a pink marker adds bean-like eyes and beaks to the figures, finalizing the bird faces.

**Unseen videos**

Figure 8: The visualization of FutureBench ID: 1q9v0XXFg6E.

1026

1027

1028 **Question:** Based on the given video, predict future events and fill in the potential events in the  
 1029 following future event slots, assuming the video ends as shown later: 1. [?] 2. [?] 3. [The man  
 1030 presents a tall, multi-layered cake with a sparkling candle].

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

**Seen videos**

1050 **A:** 1. The man performs a flip and tumbles over the countertop with flailing legs; 2. The woman appears  
 1051 concerned and cracks an egg into a bowl.

1052

1053

1054

B: 1. The woman adjusts her hair while the man sifts flour without incident; 2. Both share a quiet, calm moment  
 sorting ingredients.

1055

1056

C: 1. The man accidentally spills a bowl of chocolate leading to silence; 2. The woman scolds him and leaves  
 the kitchen.

1057

1058

D: 1. The woman glances at her phone ignoring the kitchen; 2. The man starts chopping vegetables intensely.

**Unseen videos**

Figure 9: The visualization of FutureBench ID: v\_HOTCR1uIaBM

1078

1079

1080  
 1081  
 1082  
 1083  
 1084  
 1085

**Question:** Based on the given video, predict future events and fill in the potential events in the gap before the final scene (a plated pasta dish is garnished and placed on the counter for serving). Complete the future sequence by selecting the pair of intermediate events that best bridges the observed scenes (1 to 9) to the final scene: 1. [?] 2. [?] 3. In the final scene, two white bowls of pasta are garnished by chefs with a blue finishing tool and a drizzle of green sauce, and the plated pasta is placed on a counter ready for serving.

1086  
 1087  
 1088  
 1089  
 1090  
 1091  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099  
 1100  
 1101  
 1102  
 1103  
 1104



1105  
 1106  
 1107  
 1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115

A: 1. A chef rapidly rolls out dough and cuts it into shapes using a knife; 2. Chefs exchange ingredients across the kitchen countertop before switching to dough feeding.

B: 1. Chefs feed the dough directly into a pasta machine without further preparation; 2. The pasta is immediately tossed with herbs in a bowl.

C: 1. A chef demonstrates the dough-cutting technique by first slicing the dough into precise shapes; 2. The cut dough is run through the pasta machine and its output is briefly inspected by other chefs.

D: 1. A chef sautés fresh ingredients including green herbs and vegetables in a metal pan on the stove while explaining the cooking process; 2. A chef uses a metal pan to transfer the mixture into a white bowl and then applies final plating techniques, setting the stage for the dish to be garnished in the final scene.

1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133



Figure 10: The visualization of FutureBench ID: v\_3VzXH3o88mw

1134  
1135

## C.2 QUALITATIVE ANALYSIS OF FUTURE EVENT PREDICTION TRAINING DATA

1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144

We qualitatively examine Next-Event Prediction (NEP) training pairs to verify that the future-event captions are (i) linguistically coherent and (ii) tightly grounded in the observed past video, rather than merely reflecting generic language priors. As shown in Figure 11, the past video segment shows people launching a canoe and paddling along a calm river surrounded by dense trees and cliffs. The corresponding future-event text continues this storyline: the canoe keeps moving downstream and the people begin fishing on the river. Predicting such a future requires the model to track the ongoing activity over multiple frames and infer a plausible next step in the same physical environment (continuing to paddle, then fishing), instead of describing the already-seen content or resorting to a template such as “after boating, people usually …”.

1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152

Another example is shown in Figure 12. In video part, the child approaches an arcade bowling machine, inspects the interface, and waits while the lane and pins are being prepared. In text part, the future-event caption describes the child retrieving a red ball, rolling it down the lane, and watching the pins being knocked over. Here, the future description is a concrete consequence of the earlier interaction: the child’s initial exploration of the machine, the resetting of pins, and their focused posture collectively imply that the next stage is actively playing the game. This again reflects the core NEP objective-bridging past and future segments via causal and temporal reasoning about how human actions unfold over time.

1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171

1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187

As the trip continues, two people in blue life jackets paddle a green canoe farther downstream. They move past steep, tree-covered banks, keeping a steady rhythm as the calm river reflects the surrounding greenery. The canoe drifts in and out of light current, but the pair remain in control and continue their relaxed journey. Farther along, one person in the group pauses to fish from the canoe near a rocky stretch of riverbank, casting a line toward the quieter water while others keep paddling nearby. The canoes take turns gliding through gentle ripples and returning to calmer pools, with small waves forming around the bows. More canoes join together into a loose group, with paddlers in different colored life jackets chatting and paddling at an easy pace. The river widens slightly, the sky clears, and sunlight glints off the moving water. The final shots follow the canoes from a distance as they continue downstream, cutting through the ripples and disappearing around a bend in the peaceful, forested river.

Figure 11: The visualization of training data.

1188  
 1189  
 1190  
 1191  
 1192  
 1193  
 1194  
 1195  
 1196  
 1197  
 1198  
 1199  
 1200  
 1201  
 1202 **Seen videos**  
 1203  
 1204  
 1205  
 1206  
 1207  
 1208  
 1209  
 1210  
 1211  
 1212  
 1213  
 1214  
 1215  
 1216  
 1217  
 1218  
 1219  
 1220



1221 **Future**  
 1222  
 1223  
 1224  
 1225  
 1226  
 1227

The child continues to focus on the arcade bowling game, picking up a red ball and rolling it down the lane toward the freshly reset pins. After each roll, the ball travels along the smooth lane, taps or knocks over some of the pins, and the machine briefly lights up as it clears and resets them for the next turn. The child watches each roll closely, stepping back to see the result and then moving forward again to retrieve another ball. Over several turns, the child repeats this pattern of rolling, observing the pins, and preparing for the next attempt, remaining absorbed and determined throughout the game.

1228  
 1229  
 1230  
 1231  
 1232  
 1233  
 1234  
 1235  
 1236  
 1237  
 1238  
 1239  
 1240  
 1241

Figure 12: The visualization of training data.

1242

## C.3 QUALITATIVE ANALYSIS OF MODEL PREDICTION RESULTS

1243

1244

1245

For the future prediction example in Figure 13, the predictions from Qwen2.5-VL-7B and its NEP-trained variant are shown below.

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

**Question:** Based on the given video, predict the future events and fill in the potential event in the following gap: 1. [?] 2. The person retrieves a mug from the top of the washing machine, walks towards the door, opens it, steps out of the laundry room, and closes the door behind them. Which of the following descriptions best fills the gap for the missing future event?

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

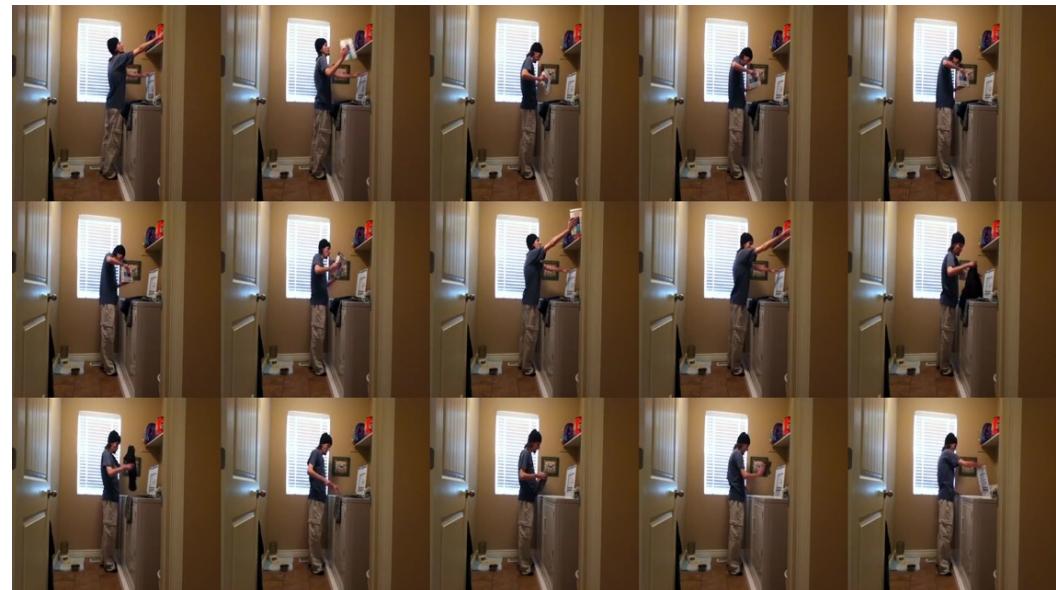
1277

1278

1279

1280

1281



1282

1283

1284

1285

1286

1287

1288

1289

1290

**A:** The person methodically sorts through the clothes on the dryer – neatly folding some items and setting others aside – and then pauses to touch their head in a moment of thoughtful decision.

**B:** The person pauses to touch their head immediately, then hurriedly shuffles the clothes from the dryer into a basket without folding.

**C:** The person directly retrieves the mug from the washing machine without addressing the pile of clothes, then walks toward the door in a rush.

**D:** The person returns to the detergent shelf to fetch another container and, while doing so, intermittently glances at the clothes on the dryer without sorting them.

1291

1292

1293

1294

1295

Figure 13: The visualization of FutureBench ID: 511AM

1296  
1297  
1298**NEP Trained Model Output for Task in Figure 13**

1299 Let's break down the video description to predict what happens next.  
 1300 The person is in a laundry room, organizing items on shelves above  
 1301 the washing machine. They handle a blue container, a white box with a  
 1302 red label, and a black garment. After placing these back, they pick  
 1303 up a dark piece of clothing from the dryer and examine it before  
 1304 putting it back. Then they take a mug from the washing machine, walk  
 1305 out, and close the door.\n\nFirst, considering the sequence of  
 1306 actions: organizing items, handling clothes, retrieving the mug. The  
 1307 person seems methodical but maybe slightly distracted, as they pause  
 1308 to touch their head. The presence of the detergent shelf suggests  
 1309 they might need more supplies, but after taking the mug, they leave.  
 1310 Since they closed the door upon exiting, perhaps they're done with  
 1311 the laundry task. However, the pile of clothes on the dryer hasn't  
 1312 been addressed yet. \n\nLooking at option A: The person is sorting  
 1313 clothes, pausing thoughtfully. This aligns with the initial  
 1314 actions-organizing and handling items. But after that, they move to  
 1315 the mug. So maybe after dealing with the mug, they might return to  
 1316 the clothes. But since they left the room, perhaps the next step is  
 1317 taking the mug elsewhere, like the kitchen. Alternatively, if they  
 1318 left the laundry room, maybe they're done, but the clothes are still  
 1319 there. However, the prediction should follow the immediate next  
 1320 steps. Since they took the mug and left, the immediate future would  
 1321 be them using the mug elsewhere, possibly in another room. But the  
 1322 question asks about future events in the video. If the video ends  
 1323 with them leaving, maybe the next scene isn't shown, but based on the  
 1324 described actions, the person has completed their tasks in the  
 1325 laundry room. Therefore, the prediction should focus on their  
 1326 departure and what they might do next with the mug. However, the  
 1327 original caption says they closed the door, so maybe the video ends  
 1328 there. But the user wants a prediction beyond the given caption.  
 1329 Wait, the caption describes the entire video up to the point where  
 1330 they exit. So the prediction is about what happens after they exit.  
 1331 But the user might want a prediction based on the last action:  
 1332 retrieving the mug and leaving. So possible outcomes: they might use  
 1333 the mug for something else, like drinking coffee, or place it  
 1334 somewhere else. Alternatively, maybe they forgot something and come  
 1335 back, but that's less likely. Given the methodical nature, perhaps  
 1336 they're done and moving on to another task. Since the video ends with  
 1337 them closing the door, the prediction could be that they proceed to  
 1338 another area, like the kitchen, to use the mug. Alternatively, if the  
 1339 mug was part of a routine, maybe they put it away. But without more  
 1340 context, the safest bet is that they exit the laundry room, having  
 1341 completed their tasks, and the next action is unrelated to the  
 1342 laundry.

1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349

1350  
1351  
1352**NEP Trained Model Output for Task in Figure 13 (Cont.)**

1353 So the answer should reflect that they've finished and are moving to  
 1354 another activity, possibly involving the mug. But since the caption  
 1355 ends with them leaving, the prediction is about what they do next.  
 1356 Maybe they take the mug to another room, like the kitchen, to have a  
 1357 drink. Alternatively, they might have forgotten something and return,  
 1358 but that's speculative. The most logical conclusion is that they  
 1359 exit, having completed their laundry-related tasks, and the mug is  
 1360 now being used for its intended purpose elsewhere. So the answer  
 1361 should mention them proceeding to another area, perhaps preparing a  
 1362 drink, using the mug, and maybe the scene transitions to show them in  
 1363 another room. But since the prediction needs to be concise, focusing  
 1364 on the immediate next step after exiting. So the answer is that they  
 1365 will likely carry the mug to another room, such as the kitchen, to  
 1366 use it, possibly indicating the end of their laundry chore and moving  
 1367 on to another task. Alternatively, they might place the mug back in  
 1368 the laundry room, but since they closed the door, that's less likely.  
 1369 Therefore, the prediction is they exit, having completed their tasks,  
 1370 and the mug is now being used elsewhere. The answer should be  
 1371 confident but allow for uncertainty if needed. Since the person  
 1372 paused to touch their head, maybe they were thinking about something  
 1373 else, but the main action is leaving with the mug. So the prediction  
 1374 is that they proceed to another area, likely to use the mug,  
 1375 completing their laundry-related activities.\nThe answer is A.

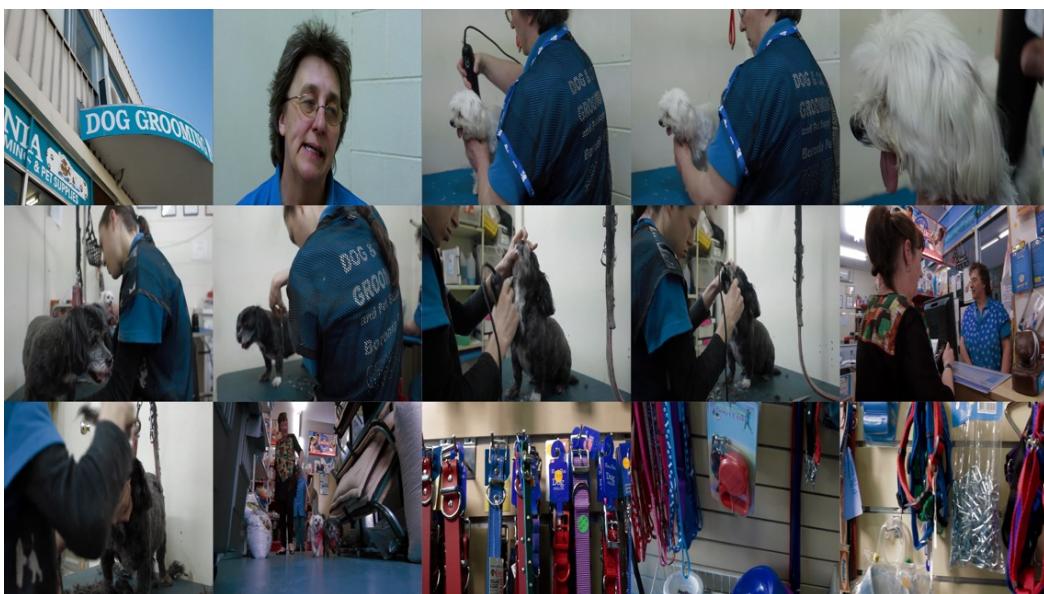
1373  
1374  
1375  
1376**Qwen2.5-VL-7B Output for Task in Figure 13**

1377 Let's analyze the sequence of actions in the video and the options  
 1378 provided:\n\n1. The person is initially seen reaching up to the shelf  
 1379 above the washing machine.\n2. They retrieve several items (such as  
 1380 detergent and a container), briefly inspect them, and place them back  
 1381 on the shelf.\n3. They then move to the dryer, pick up a dark piece  
 1382 of clothing, look at it for a moment, and put it back without folding  
 1383 or sorting the rest of the pile.\n4. Instead of continuing with the  
 1384 laundry, they shift their attention to the washing machine, retrieve  
 1385 a mug from inside, and immediately head toward the door.\n5. They  
 1386 walk out of the laundry room with the mug and close the door behind  
 1387 them, leaving the pile of clothes on the dryer unaddressed.\n\nNow,  
 1388 let's evaluate each option:\n\nA. This option describes the person  
 1389 methodically sorting through the clothes on the dryer and pausing  
 1390 thoughtfully. While the person does briefly handle a piece of  
 1391 clothing, there is no clear indication that they engage in a  
 1392 sustained or systematic sorting process, so this does not fully match  
 1393 the observed sequence.\nB. This option suggests that the person  
 1394 pauses to touch their head immediately and then hurriedly shuffles  
 1395 the clothes into a basket without folding them. This rushed,  
 1396 bulk-moving behavior is not supported by what we see in the  
 1397 video.\nC. This option states that the person retrieves the mug from  
 1398 the washing machine without properly dealing with the pile of clothes  
 1399 and walks toward the door to leave, which closely matches the  
 1400 observed behavior: they only briefly interact with the clothes and  
 1401 then prioritize taking the mug and exiting the room.\nD. This option  
 1402 mentions that the person returns to the detergent shelf to fetch  
 1403 another container and keeps glancing at the clothes without really  
 1404 acting on them, which does not align with the fact that they leave  
 1405 the room right after taking the mug.\n\nGiven this analysis, the most  
 1406 accurate description of the future events is option C.\n\nThe answer  
 1407 is C.

1403

1404  
 1405 Another example is shown in Figure 14. In the dog-grooming video, the base model (without NEP)  
 1406 prefers option A, which inserts yet another close-up grooming scene. This is textually plausible given  
 1407 the prompt (“dog grooming & pet supplies”) and reflects a strong prior that such videos should keep  
 1408 showing grooming actions, but it ignores the **temporal role** of the missing clips: they are supposed  
 1409 to be the last interior segments immediately before the closing exterior storefront shot, after multiple  
 1410 grooming interactions have already been shown. In contrast, the NEP-trained model selects option  
 1411 B, which focuses on a specific pet bed with brochures followed by a wide product-shelf shot. This  
 1412 sequence is less “obvious” from language priors alone, but it matches the actual chronology and  
 1413 structure of the advertisement: the video first establishes the service (grooming), then transitions  
 1414 to product displays, and only then returns to the exterior sign for branding. The NEP model thus  
 1415 demonstrates sensitivity to temporal structure and previously observed content (it avoids re-describing  
 1416 past grooming events and chooses scenes that function as a bridge to the final shot), rather than simply  
 1417 following generic textual expectations about what typically happens in a dog-grooming video.  
 1418

1419 **Question:** Based on the given video, predict future events and fill in the potential events in the following  
 1420 sequence: 1. [?] 2. [?] 3. [A final exterior shot of the building with the 'DOG GROOMING' sign on a teal  
 1421 background]. Which option correctly represents the intermediate scenes that would logically occur  
 1422 between the observed scenes and the final scene?



1423  
 1424 A: A detailed close-up of colorful dog collars displayed on a rack with price tags; then a transition to a person in a  
 1425 blue shirt demonstrating grooming on a small white dog using electric clippers in an indoor setting.  
 1426  
 1427 **B: An expansive product display featuring a pet bed with embedded brochures promoting Flea-Free Dog Bed**  
 1428 **and MultiMat; followed by a comprehensive shelf scene showcasing an array of pet supplies including dog**  
 1429 **toys, grooming tools, and food packages.**  
 1430  
 1431 C: A counter conversation between a person with dark hair in a ponytail and another staff member inside a store;  
 1432 then a quick montage of various grooming tools laid out on a table.  
 1433  
 1434 D: An outdoor shot panning across different parts of the building's exterior; then a brief cut to a montage of pet  
 1435 supplies being stocked on shelves.  
 1436  
 1437

1438  
 1439  
 1440 **Figure 14: The visualization of FutureBench ID: v1-3-train\_val-v\_J6T9pokWX1E**  
 1441  
 1442  
 1443  
 1444  
 1445  
 1446  
 1447  
 1448  
 1449

1458

**NEP Trained Model Output for Task in Figure 14**

1459

1460

1461 Let's start by breaking down the video provided. The video begins  
 1462 with the exterior of Boronia Dog Grooming & Pet Supplies,  
 1463 establishing the location. Then it moves indoors to show staff  
 1464 members grooming dogs, highlighting their uniforms with the company  
 1465 name. There are scenes of different dogs being groomed, interactions  
 1466 between staff, and displays of pet products like collars, leashes,  
 1467 and beds. The final scene mentioned is an exterior shot of the  
 1468 building with the 'DOG GROOMING' sign.\n\nTo predict future events  
 1469 after the final scene described, I need to consider the flow of the  
 1470 video. The video seems to be promotional, showcasing the services and  
 1471 products of the grooming salon. Typically, such videos might end with  
 1472 contact information, a call to action, or a return to the exterior to  
 1473 reinforce the business details. Since the last scene is an exterior  
 1474 shot, the next logical step could be displaying the business hours,  
 1475 phone number, or website. Alternatively, they might show happy pets  
 1476 and satisfied customers leaving, but since the final scene is already  
 1477 exterior, maybe adding text overlays with key info.\n\nLooking at the  
 1478 existing structure: exterior to interior staff grooming to product  
 1479 displays to exterior again. The video might loop back to the exterior  
 1480 with additional info. The initial exterior had a sign with 'DOG  
 1481 GROOMING', so perhaps adding the full address, phone number, or  
 1482 opening hours. Alternatively, a closing shot with the logo and  
 1483 contact details superimposed.\n\nBut the user wants future  
 1484 predictions based on the current scenes, not necessarily the very  
 1485 end. However, the final scene described is the exterior shot. So  
 1486 maybe the prediction should focus on what happens immediately before  
 1487 that final shot. Wait, the caption says the video concludes with that  
 1488 exterior shot. So the future prediction would be after that? But the  
 1489 task is to predict future events based on the scene, so maybe the  
 1490 user expects the prediction to be about what comes after the  
 1491 described scenes, but according to the caption, the video ends with  
 1492 the exterior shot. Hmm, this is a bit confusing.\n\nWait, the user's  
 1493 instruction says: \"Based on the scene, analyze and predict future  
 1494 events. Provide concise, natural, and confident prediction about the  
 1495 future events of the video.\" The caption describes the entire video  
 1496 up to the conclusion. So perhaps the user wants the assistant to  
 1497 imagine that the video continues beyond the described scenes. But  
 1498 according to the caption, the video as described ends with the  
 1499 exterior shot. Therefore, the prediction should be about what would  
 1500 happen next if the video continued beyond that  
 1501 point.\n\nAlternatively, maybe the user made a mistake in the  
 1502 caption, thinking that the video continues, but according to the  
 1503 description, it's concluded. This is a bit ambiguous.

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515 However, given the task, we have to work with the information  
 1516 provided.\n\nAssuming the video is partway through and the final  
 1517 scene is yet to come, but according to the caption, the video has  
 1518 already concluded with that exterior shot. Therefore, perhaps the  
 1519 user intended the final scene to be part of the video, and the  
 1520 prediction is about what would follow naturally after the scenes  
 1521 described. For example, after showing the exterior, maybe  
 1522 transitioning to contact details or a logo.\n\nBut the user's  
 1523 instruction says to predict future events based on the scene. If the  
 1524 video as described ends with the exterior shot, then there are no  
 1525 more scenes to predict. However, maybe the user expects the assistant  
 1526 to infer that after showcasing the grooming and products, the next  
 1527 logical step is to display contact information or a call to action.  
 1528 Alternatively, the video might show a satisfied customer or a dog  
 1529 leaving the salon, but since the final scene is exterior, perhaps  
 1530 adding text with the business details.\n\nAlternatively, considering  
 1531 the structure of promotional videos, after showing the products and  
 1532 services, they often end with contact details. So the prediction  
 1533 could be that the video will display the business's contact  
 1534 information (phone number, website) over the exterior shot, or  
 1535 transition to a logo with the tagline.\n\nBut since the user wants  
 1536 the prediction to be based on the scenes described, which include the  
 1537 exterior, interior grooming, product displays, and staff  
 1538 interactions, the next logical step after these scenes would be to  
 1539 wrap up with contact info. However, the final scene described is the  
 1540 exterior shot. Therefore, the prediction should be about what would  
 1541 come after that, but according to the caption, the video ends there.  
 1542 Therefore, perhaps the assistant should infer that the video would  
 1543 conclude with contact details superimposed on the exterior  
 1544 shot.\n\nAlternatively, maybe the video is structured to show the  
 1545 exterior, then move into the interior, then products, then staff,  
 1546 then back to exterior with contact info. But the caption says the  
 1547 video concludes with the exterior shot. So perhaps the prediction is  
 1548 that after the exterior shot, the video might display contact  
 1549 information for Boronia Dog Grooming, such as a phone number or  
 1550 website, possibly with a voiceover or text overlay.\n\nBut the user  
 1551 wants the prediction to be concise, natural, and confident, avoiding  
 1552 references to reading captions. So phrasing it as observing the  
 1553 scene, like: \"The video will likely conclude by displaying contact  
 1554 details for Boronia Dog Grooming, such as a phone number or website,  
 1555 overlaid on the exterior shot to encourage viewers to  
 1556 visit.\"\\n\nHowever, the user's example response uses options from a  
 1557 multiple-choice question. The answer was \"The answer is B\".\n

1558

1559

1560

1561

1562

1563

1564

1565

1566  
1567  
1568**Qwen2.5-VL-7B Output for Task in Figure 14**

1569 To determine the correct sequence of events that logically follows  
 1570 the given video clips, let's analyze the content of each clip and the  
 1571 options provided:  
 1572 The first clip shows the exterior of  
 1573 `"BORONIA DOG GROOMING & PET SUPPLIES"` with a focus on the 'DOG  
 1574 GROOMING' sign.  
 1575 The second clip transitions to a woman in a blue  
 1576 shirt grooming a small white dog with electric clippers.  
 1577 The  
 1578 third clip shifts to a woman in a black shirt interacting with a  
 1579 customer at a counter.  
 1580 The fourth clip shows a woman grooming a  
 1581 black dog with clippers.  
 1582 The fifth clip displays a variety of dog  
 1583 collars and leashes on a rack.  
 1584 The sixth clip showcases a pet bed  
 1585 with brochures for 'Flea-Free Dog Bed' and 'Multimat'.  
 1586 The  
 1587 seventh clip presents a shelf with various pet supplies, including  
 1588 dog toys, grooming tools, and food packages.  
 1589 The eighth clip  
 1590 shows a woman grooming a small white dog with clippers.  
 1591 The ninth  
 1592 clip features a woman grooming a black dog with clippers.  
 1593 The tenth clip displays a variety of dog collars and leashes on a  
 1594 rack.  
 1595 The eleventh clip shows a counter conversation between two  
 1596 people.  
 1597 The twelfth clip presents a quick montage of grooming  
 1598 tools on a table.  
 1599 The final clip shows a final exterior shot of  
 1600 the building with the 'DOG GROOMING' sign on a teal  
 1601 background.  
 1602 Now, let's evaluate the options:  
 1603 **A.** This option  
 1604 starts with a close-up of dog collars, which fits the context of the  
 1605 video as it shows pet supplies. It then transitions to a grooming  
 1606 session, which also aligns with the video content. This option seems  
 1607 logical and follows the sequence well.  
 1608 **B.** This option starts with a product display of a pet bed, which is not directly related to the  
 1609 initial exterior shot of the grooming salon. It then moves to a shelf  
 1610 scene, which is also not directly connected to the initial grooming  
 1611 session. This option does not follow the sequence logically.  
 1612 **C.** This option starts with a counter conversation, which is not directly  
 1613 related to the initial exterior shot of the grooming salon. It then  
 1614 moves to a montage of grooming tools, which is more aligned with the  
 1615 video content but does not follow the sequence logically from the  
 1616 initial exterior shot.  
 1617 **D.** This option starts with an outdoor shot  
 1618 of the building, which is similar to the initial exterior shot but  
 1619 does not provide any direct connection to the grooming session. It  
 1620 then moves to a montage of pet supplies, which is not directly  
 1621 related to the initial exterior shot.  
 1622 Given the analysis, the most  
 1623 logical sequence that follows the initial exterior shot and leads to  
 1624 the final exterior shot is Option A.  
 1625 The answer is A.

1603

1604 In the following, we discuss the failure pattern of the NEP-trained model. As shown in Figure 15 and  
 1605 Figure 16.

1606

1607 As the model response for example in Figure 15, the model overlooks the two explicit future anchors  
 1608 specified in the textual question and fails to precisely localize the step to the stage where ingredients  
 1609 are still being incrementally added. Instead, it relies on generic cocktail-making knowledge and treats  
 1610 "shaking and plating the bottle" as a plausible subsequent action, thus incorrectly choosing option B.

1611

1612 Similarly, in the hair-styling example (shown in Figure 16), the model again produces a  
 1613 reasonable future continuation at the narrative level—it imagines the person continuing to section the  
 1614 hair, re-apply oil, and use the spray bottle for final touches—but fails to align this prediction with the  
 1615 way the question specifies future events. Instead of leveraging the discriminative visual anchors that  
 1616 actually appear later in the video (the yellow hair clip and the black "Sunsilk" tube with blue accents),  
 1617 it defaults to a generic hair-tutorial script and therefore selects option B rather than the ground-truth  
 1618 option C, which more faithfully reflects the true future frames.

1619

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631

1632 Question: Based on the given video, predict future events following the observed steps. Two key future  
1633 moments are provided: one where the person is seen picking up a large bottle of Ciroc Apple vodka and  
1634 pouring it into the Smirnoff Ice bottle with the decorative candy strip still on it, and later, a scene where  
1635 additional Ciroc Apple vodka is poured into the bottle, followed by a red sauce from a yellow bottle.  
1636 Which intermediate event is most likely to occur between these two moments?



1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

**A:** The person adds chopped fruits, including pieces of apple and strawberry, into the bottle.  
**B:** The person shakes the bottle and places it on a black cloth.  
**C:** The person wraps a rainbow-colored candy strip around the neck of the bottle.  
**D:** The person uses a wooden skewer to assemble a garnish made of various colorful gummy candies.

Figure 15: The visualization of FutureBench ID: `wwNaAxypK4o`

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

**NEP Trained Model Output for Task in Figure 15**

Let's break down the video step by step to predict what might happen next. The person starts by preparing a Smirnoff Ice bottle, coating it with dark brown liquid (likely melted chocolate) using a spoon. Then they dip the bottle into a bowl of white powder (probably sugar or salt), creating a textured coating. After that, they wrap a colorful rainbow candy strip around the bottle, securing it with a toothpick. \n\nNext, they place the decorated bottle on a white surface and pour clear liquid (Smirnoff Ice) into a glass with ice cubes. Then there's a transition to them holding a large Ciroc Apple bottle, pouring its contents into the Smirnoff Ice bottle. The process continues with adding more Ciroc Apple and a red sauce from a yellow bottle.\n\nNow, considering the sequence, the person is likely creating a layered cocktail or mixed drink. The decorative elements suggest a focus on presentation, possibly for a social media tutorial or a creative recipe showcase. The use of different vodkas (Smirnoff Ice and Ciroc Apple) and the addition of a red sauce (maybe something like strawberry syrup or hot sauce?) indicates mixing flavors.\n\nAfter pouring the Ciroc and the red sauce into the Smirnoff Ice bottle, the next logical step would be to seal or close the bottle. Since they used a toothpick to secure the candy strip, maybe they'll add another element to keep it closed, like a lid or a decorative cap. Alternatively, they might shake the bottle to mix the ingredients, which would be a common step in cocktail making. However, since the bottle is already filled with layers, shaking could mix everything, but if they want to preserve the layers, they might not shake it. Instead, they could present the finished product, perhaps placing it back on the counter or adding a garnish on top.\n\nLooking at the previous steps involving meticulous decoration (chocolate coating, sugar/salt, candy strip), the next action might involve adding a final touch, like sprinkling more toppings, drizzling something else, or inserting a skewer with colorful candies as a garnish. The presence of the red sauce suggests a spicy or fruity element, so maybe they'll add a complementary garnish, like a strawberry slice or apple piece, especially since Ciroc Apple is mentioned.

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

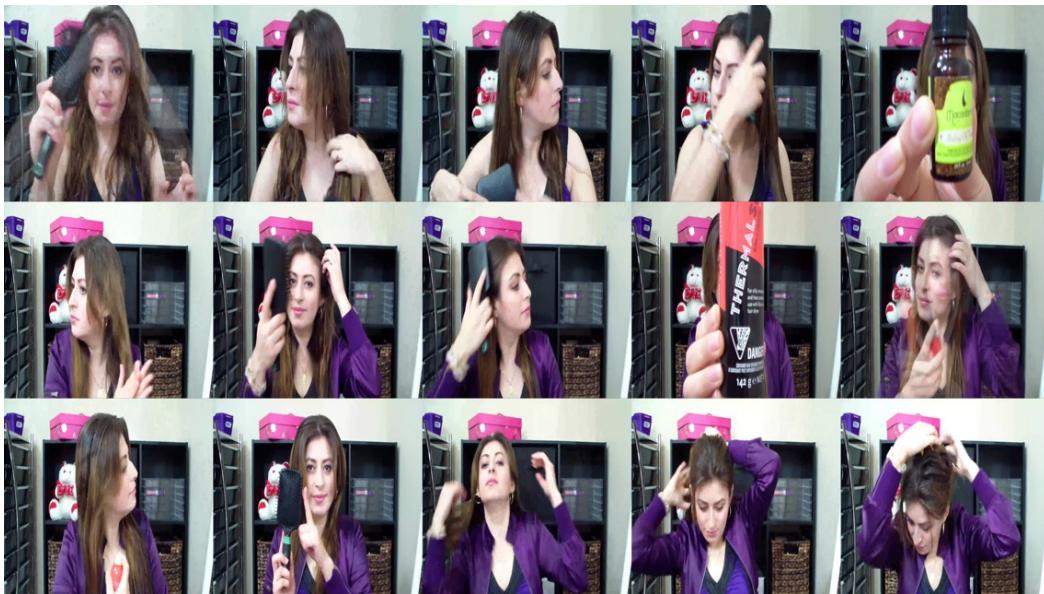
1781

**NEP Trained Model Output for Task in Figure 15 (Cont.)**

Alternatively, they might use the wooden skewer again to insert a gummy candy or other decorative item.\n\nAnother possibility is that after filling the bottle, they might demonstrate pouring the mixture into glasses, but since the current scene shows pouring into a glass with ice earlier, maybe they're building a layered drink inside the Smirnoff Ice bottle itself. If that's the case, the next step could be adding another ingredient on top, like a foam or another liquid layer, but the video doesn't mention other liquids beyond the Ciroc and red sauce. Alternatively, they might present the finished bottle as a centerpiece, maybe placing it in a visually appealing setup with other drinks or snacks.\n\nConsidering the transitions in the video described, each step builds on the previous one with increasing decoration and mixing. The final steps would likely emphasize presentation, so the person might place the bottle on a tray with other items, add a straw or a decorative pick, or even take a sip to taste. However, since the last action was pouring the red sauce, the immediate next step could be sealing the bottle, perhaps using the toothpick again to attach a lid or a small umbrella, or maybe adding a final sprinkle of the white powder (like powdered sugar) for visual contrast.\n\nGiven that the person has been meticulous in their decorations, the prediction should align with continuing that trend. Therefore, the most probable next actions are either sealing the bottle with a decorative element or adding a final garnish. Since they've already secured the candy strip with a toothpick, maybe they'll add another layer of decoration on top, such as drizzling chocolate or sprinkling herbs, but the red sauce might be part of the flavor profile rather than decoration. Alternatively, they might present the bottle as completed, perhaps placing it in a display or adding a label with the recipe name.\n\nAnother angle is considering the tools used: they have a wooden skewer, toothpick, and bowls for coating. They might use the skewer again to insert more candies or create a stand for the bottle. But since the bottle is upright and filled, maybe they'll add a skewer through the cap to hold a garnish.\n\nIn summary, based on the progression from decorating the bottle to adding multiple liquids, the next logical step is either sealing the bottle with a final touch or presenting the finished product. Given the prior use of toothpicks and candy strips, adding a final decorative element seems plausible. So, the prediction would be that the person will secure the bottle with a lid or a decorative element, possibly using the toothpick again, and then present the finished drink, maybe placing it alongside other items or demonstrating consumption.\n\nThe answer is B.

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791

1792 **Question:** Based on the given video, predict future events and fill in the potential events in the following future  
1793 sequence: 1. [?] 2. [?] 3. The video concludes with the person adjusting their hair, smoothing it down, and  
1794 showcasing its final smooth and shiny texture. Which option best describes the intermediate events that occur after  
1795 the observed scenes?



1806 A: 1. The person gathers their hair into a ponytail and styles it using a black hair dryer with a white round brush; 2.  
1807 The person applies an extra coat of product to the hair.  
1808 B: 1. The person sectioned the hair using additional clips and re-applied Moroccan Healing Oil; 2. The person uses  
1809 the red spray bottle for final touches on the hair.",  
1810 C: 1. The person gathers their hair into a ponytail using a yellow hair clip and dries and styles it with a black  
1811 hair dryer and a white round brush; 2. The person holds up a black tube labeled 'Sunsilk' with blue accents,  
1812 examines it, and then puts it down.  
1813 D: 1. The person repositions the white teddy bear on the shelf while spattering thermal spray; 2. The person uses a  
1814 new technique of running fingers through the hair, creating loose curls.

1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

Figure 16: The visualization of FutureBench ID: v\_RkCSfz8GeNg

1836  
1837  
1838**NEP Trained Model Output for Task in Figure 16**

1839 Let's break down the video step by step to predict what comes next.  
 1840 The person starts by showing their hair routine, using specific  
 1841 products like the Moroccan Oil and Thermal Spray. They sectioned  
 1842 their hair, applied products, used a brush, and then clipped their  
 1843 hair up. The key actions here are methodical application of each  
 1844 product, sectioning, and styling.  
 1845 After clipping their hair up, the next logical step would be to either continue drying another  
 1846 section or apply more products. Since they already used the Thermal  
 1847 Spray and the black tube (possibly a heat protectant), maybe they'll  
 1848 move to another part of the hair. The mention of gathering into a  
 1849 ponytail suggests they might be preparing to style it further. The  
 1850 presence of the hair dryer and brushes indicates they might start  
 1851 blow-drying the clipped sections, perhaps using the round brush for  
 1852 volume.  
 1853 Also, since they showed the products to the camera, they  
 1854 might demonstrate the application on the clipped sections, maybe  
 1855 applying more oil or thermal spray. The final steps would likely  
 1856 involve using the hair dryer to dry the clipped hair, then brushing  
 1857 it out for volume. The ponytail might be styled into a sleek look or  
 1858 incorporated into the routine for added texture.  
 1859 Considering the flow, after clipping, they might take a section from the ponytail  
 1860 area, apply product, and use the brush and dryer to smooth or add  
 1861 shine. Then repeat with other sections until all hair is styled. The  
 1862 conclusion would show the final look, as mentioned in the caption. So  
 1863 the intermediate steps would involve continuing the process on the  
 1864 remaining hair, possibly using the same products again, and then  
 1865 finishing with the dryer and brush to achieve the smooth texture  
 1866 shown at the end.  
 1867 Looking at the options provided, the one that  
 1868 aligns with these steps is B. It mentions sectioning with clips,  
 1869 reapplying oil, and using the spray bottle for final touches. That  
 1870 fits the flow. Option A has them gathering into a ponytail first,  
 1871 which might come later, but the current step is clipping. Option C  
 1872 involves a different clip color and a different product label, which  
 1873 doesn't match. Option D talks about repositioning the teddy bear,  
 1874 which seems unrelated. So B is the most plausible.  
 1875 Therefore, the  
 1876 prediction should outline the continuation of styling the clipped  
 1877 sections, applying products, using the tools shown, leading to the  
 1878 final smooth result. The answer is B.  
 1879 The answer is B.

1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889

1890  
1891

## C.4 THREE TYPES OF LOGIC REASONING IN VIDEO INSTRUCTION TUNING

1892

- Induction Video Prediction (Video Q&A):  
“<Full Video> When is the best time to add the meat learned from the video?”



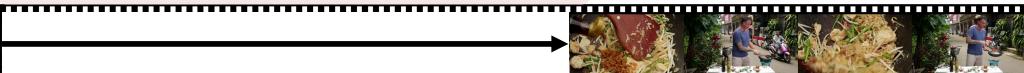
1897

- Deduction Video Prediction (Next Event Prediction):  
“<First Part Video> What will happen next?”



1902

- Abduction Video Prediction (Previous Event Prediction):  
“<Second Part Video> What happened beforehand?”



1906

Figure 17: **Three types of logic reasoning in video instruction tuning tasks.** (1) **Induction** (Video Q&A): The model watches entire video sequences and learns common event patterns and temporal relationships, building an internal “engine” of how visual events unfold over time. (2) **Deduction (Next Event Prediction)**: Given the first part of a video, the model uses its learned causal and commonsense knowledge to extrapolate and predict the most likely next events. (3) **Abduction** (Previous Event Prediction): Presented with the final segment of a video, the model reasons backward to hypothesize plausible prior events or hidden causes that explain the observed outcome.

1913

1914

1915

## D TRAINING STRATEGY

1916

**Supervised Fine Tuning (SFT).** We fine-tune the MLLM on V1-33K using standard supervised learning. The model receives the first segment of a video caption and predicts its continuation, training via cross-entropy loss. This stage instills basic predictive capabilities, allowing the model to directly imitate ground-truth future event descriptions.

1921

**Critique Fine Tuning (CFT).** CFT is a strategy where models learn to critique noisy responses instead of simply imitate answers Wang et al. (2025). We leverage critique data generated by an external LLM (e.g., GPT-4) that identify strengths and errors in model predictions relative to ground-truth continuations. During fine-tuning, the model learns to refine flawed continuations or evaluate predictions based on provided critiques, internalizing feedback to enhance logical consistency and predictive accuracy.

1928

**Distillation Tuning (Distill).** We employ knowledge distillation from DeepSeek-R1, a strong reasoning model. For each sample, DeepSeek-R1 generates detailed reasoning steps and a predicted caption. The student model is fine-tuned to reproduce this entire reasoning sequence, adopting structured inferential patterns to improve both reasoning and prediction accuracy.

1932

**Mix Tuning (Mix).** We combine SFT, CFT, and Distillation methods equally in each training epoch. By interleaving direct predictions, critique-informed refinements, and explicit reasoning demonstrations, the model integrates various supervision signals. This mixed strategy promotes robust learning, balancing factual accuracy, critical feedback integration, and structured reasoning capabilities.

1938

1939

## E PROMPTS

1940

## E.1 EVENT IDENTIFICATION PROMPT

1941

This prompt ensures structured extraction of discrete events from raw captions.

1944  
1945  
1946  
1947     Below is the video caption:  
1948     {video\_caption}  
1949     Task:  
1950     1. Identify and list the events (scenes) in the video in sequential  
1951         order (e.g., Scene 1, Scene 2, etc.).  
1952     2. For each scene, provide a description.  
1953  
1954     Please return your answer in a valid JSON format exactly as follows  
1955         (with no extra text):  
1956  
1957     {  
1958         "events": [  
1959         {"scene": "Scene 1",  
1960             "description": "Brief description of scene 1"},  
1961         {"scene": "Scene 2",  
1962             "description": "Brief description of scene 2"},  
1963         ...  
1964         ]  
1965     }

## E.2 CAUSAL ANALYSIS AND SPLITTING SUITABILITY PROMPT

1966  
1967  
1968     This prompt analyzes the causal dynamics among events and determines an optimal split point that  
1969         yields a stronger supervision signal for future prediction.

### Causal Analysis and Splitting Suitability Prompt

1970  
1971  
1972  
1973     Below are the extracted events from the video:  
1974     {json.dumps(event\_identification\_result, indent=2)}  
1975  
1976     Original video caption:  
1977     {video\_caption}  
1978     Task:  
1979     1. Analyze the causal relationships among these events.  
1980     2. Determine whether the video is suitable to be split into two parts  
1981         for causal inference (i.e., given the first part, can we predict  
1982         what happens in the second part?).  
1983     3. If it is suitable, specify the optimal split point (for example,  
1984         'between Scene A and Scene B').  
1985  
1986     Please provide your answer in a valid JSON format exactly as follows  
1987         (with no extra text):  
1988  
1989     {  
1990         "suitable": "yes" or "no",  
1991         "optimal\_split\_point":  
1992             "between Scene X and Scene Y",  
1993         "reasoning":  
1994             "Detailed explanation of the causal relationships  
1995             and the split decision."  
1996     }  
1997

1998  
1999

## E.3 CAPTION SPLITTING PROMPT

2000

This prompt divides the caption into meaningful segments at the identified split.

2001

**Caption Splitting Prompt**

2003

2004

Using the identified events and the optimal split point, split the original video caption into two parts. The optimal split point is given in the format 'between Scene X and Scene Y'. This means that all scenes up to and including Scene X should be included in the first part ('caption\_part1'), and all scenes from Scene Y onward should be included in the second part ('caption\_part2').

2009

2010

The identified events:

2011

```
{json.dumps(event_identification_result, indent=2)}
```

2012

and the optimal split point:

2013

```
{casual_analysis_result["optimal_split_point"]}
```

2014

2015

Original video caption:

2016

2017

```
{video_caption}
```

2018

2019

Return your answer in a valid JSON format exactly as follows (no extra text):

2020

```
{
    "caption_part1": "Text for first part",
    "caption_part2": "Text for second part"
}
```

2024

2025

2026

2027

## E.4 CHAIN-OF-THOUGHT REASONING &amp; FUTURE PREDICTION PROMPT

2028

2029

This prompt guides the reasoning model to forecast upcoming events, and the resulting reasoning trace is then used for training.

2030

2031

**Chain-of-Thought Reasoning & Future Prediction Prompt**

2032

2033

2034

You have advanced visual perception abilities and can analyze videos as if you are watching them in real time. You will be provided with a detailed description of a video (caption). Interpret this description as if it represents your actual dynamic visual experience rather than just text.

2035

2036

Based on the scene, analyze and predict future events. Provide concise, natural, and confident prediction about the video's future events. Speak as if you are directly observing the events, avoiding any reference to reading text or captions. If details are ambiguous, express natural uncertainty (e.g., "It appears that ...").

2037

2038

Caption:

2044

```
{caption_part1}
```

2045

2046

2047

2048

2049

2050

2051

2052 E.5 FUTURE PREDICTION CRITIQUE PROMPT  
20532054 This prompt critically evaluates the alignment of predictions with the actual video outcome.  
20552056 Future Prediction Critique Prompt  
2057

2058 Task:

2059 Review the caption of the second part of a video as the ground truth  
2060 and evaluate whether the future prediction (derived from the first  
2061 part of the video) aligns with the actual events.2062 What actually happened in the second part of the video:  
2063

2064 {caption\_part2}

2065

2066 Prediction (derived from the first part of the video):  
2067

2068 {prediction\_content}

2069 Reasoning behind the prediction:  
2070

2071 {reasoning\_content}

2072 Instructions:

2073 1. Analyze the prediction and the reasoning provided, considering how  
2074 well they align with the ground truth.  
2075 2. Note that accurately predicting future events is inherently  
2076 challenging; allow for minor discrepancies and avoid overly strict  
2077 judgments.  
2078 3. Think step by step and provide a critique of the prediction and  
2079 its underlying reasoning.  
2080 4. Conclude your analysis by stating either "Conclusion: right" if  
2081 the prediction aligns well, or "Conclusion: wrong" if it does not.

2082 Output:

2083 Return your analysis in a valid JSON format exactly as shown below  
(do not include any extra text):  
20842085 {  
2086     "Critique":  
2087         "Your critique of the prediction and its underlying reasoning",  
2088         "Conclusion": "right"/"wrong"  
2089 }  
20902091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105

2106 E.6 REWRITE PROMPT  
21072108 The rewriting prompt serves as a data-cleaning step to further refine the wording of the generated  
2109 reasoning traces.2110 **Rewrite Reasoning Prompt**  
21112112  
2113 You will receive a snippet of text that references a "description" or  
2114 "caption" of a video. Your task is to produce a **nearly identical**  
2115 version of that text with **minimal** changes, focusing on the  
2116 following:2117  
2118 1. **Replace references to "description" or "caption"** with wording  
2119 that references **"the video."**  
2120 - For example, "The description says..." could become  
2121 "The video shows..."  
2122 - "The caption suggests..." could become  
2123 "The video suggests..."  
2124 - Make sure the replacement sounds natural but does  
2125 **not** otherwise change the meaning.  
2126  
2127 2. **Preserve all line breaks, punctuation, and spacing** as much as  
2128 possible, and make **no additional edits** outside of these  
2129 replacements.  
2130  
2131 3. You should only output the rewritten content.2132  
2133 Here is the input:  
2134 {reasoning\_content}2135 **Rewrite Prediction Prompt**2136 You will receive a snippet of text that references a "description" or  
2137 "caption" of a video. Your task is to produce a **nearly identical**  
2138 version of that text with **minimal** changes, focusing on the  
2139 following:2140  
2141 1. **Replace references to "description" or "caption"** with wording  
2142 that references **"the video."**  
2143 - For example, "The description says..." could become  
2144 "The video shows..."  
2145 - "The caption suggests..." could become  
2146 "The video suggests..."  
2147 - Make sure the replacement sounds natural but does  
2148 **not** otherwise change the meaning.2149 Here is the input:  
2150 {prediction\_content}2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

2160 E.7 FUTUREBENCH CONSTRUCTION PROMPT

2161

2162 **FutureBench 1-Hop Question Construction Prompt**2163 *This prompt aims to generate the 1-hop QA pairs of FutureBench.*

2164

2165 **FutureBench 1-Hop Question Construction Prompt**

2166

2167 You are an expert in video understanding. Your task is to generate  
2168 one multiple-choice question to assess the video understanding  
2169 ability of a test model. You are given the meta information about a  
2170 video that includes:

- 2171 - Video captions: A complete description of the entire video for your  
reference.
- 2172 - Scene descriptions: Detailed descriptions of key scenes throughout  
the video.
- 2173 - Observed Scenes: Scenes in the given video that the test model can  
observe.
- 2174 - Last Scene: The last scene of the entire video.

2175

Requirements:

2176

## 1. Question Content:

- 2177 - Given the video with observed scenes (scene 1 to k), the question  
should force the test model to predict future events (scene k+1 to  
2178 scene n) and ask what intermediate events would be supposing scene n  
is given and scene n is the potential future end.
- 2179 - For example, "Question": "Based on the given video, predict future  
events and fill in the potential events in the given future events:  
2180 1. [?] 2. [describe scene n]. "Options": A/B/C/D. [describe scene for  
slot 1]
- 2181 - Keep the event slot [?] to be filled.
- 2182 - Construct the future event gap so that it is hard enough. For  
example, wrong answers could present the wrong order of the predicted  
2183 future events.
- 2184 - Avoid using scene id in the question and start the question from  
2185 "Based on the given video, ..."

2186

## 2. Question Format:

- 2187 - Create one multiple-choice question with four answer options: A, B,  
2188 C, and D.
- 2189 - Ensure only one correct answer and that the remaining three options  
are wrong.
- 2190 - Only output required question-answer pairs shown in the output  
structure.

2191

Output structure:

2192

{output\_structure}

2193

Please generate an example question based on the following input data.

2194

## Input Data:

- 2195 - Video captions: {caption}
- 2196 - Scene descriptions: {event}
- 2197 - Observed Scenes: {obs}
- 2198 - Last Scene: {last}

2199

2200

2201

2202

2203

2204

2205

2206

2207

2208

2209

2210

2211

2212

2213

2214

**FutureBench 2-Hop Question Construction Prompt**

2215

*This prompt aims to generate the 2-hop QA pairs of FutureBench.*

2216

2217

**FutureBench 2-Hop Question Construction Prompt**

2218

2219

2220

2221

2222

2223

2224

2225

2226

2227

2228

2229

2230

2231

2232

2233

2234

2235

2236

2237

2238

2239

2240

2241

2242

2243

2244

2245

2246

2247

2248

2249

2250

2251

2252

2253

2254

2255

2256

2257

2258

2259

2260

2261

2262

2263

2264

2265

2266

2267

You are an expert in video understanding. Your task is to generate one multiple-choice question to assess the video understanding ability of a test model. You are given the meta information about a video that includes:

- Video captions: A complete description of the entire video for your reference.
- Scene descriptions: Detailed descriptions of key scenes throughout the video.
- Observed Scenes: Scenes in the given video that the test model can observe.
- Last Scene: The last scene of the entire video.

Requirements:

1. Question Content:

- Given the video with observed scenes (scene 1 to k), the question should force the test model to predict future events (scene k+1 to scene n) and ask what intermediate events would be supposing scene n is given and scene n is the potential future end.
- For example, "Question": "Based on the given video, predict future events and fill in the potential events in the given future events: 1. [?] 2. [?] 3. [describe scene n]. "Options": A/B/C/D. [describe scene for slot 1], [describe scene for slot 2]
- Keep the event slot [?] to be filled.
- Construct the future event gap so that it is hard enough. For example, wrong answers could present the wrong order of the predicted future events.
- Avoid using scene id in the question and start the question from "Based on the given video, ..."

2. Question Format:

- Create one multiple-choice question with four answer options: A, B, C, and D.
- Ensure only one correct answer and that the remaining three options are wrong.
- Only output required question-answer pairs shown in the output structure.

Output structure:

{output\_structure}

Please generate an example question based on the following input data.

Input Data:

- Video captions: {caption}
- Scene descriptions: {event}
- Observed Scenes: {obs}
- Last Scene: {last}

2268  
2269  
2270**FutureBench 3-Hop Question Construction Prompt**  
*This prompt aims to generate the 3-hop QA pairs of FutureBench.*2271  
22722273  
2274**FutureBench 3-Hop Question Construction Prompt**

2275

2276  
2277  
2278  
2279

You are an expert in video understanding. Your task is to generate one multiple-choice question to assess the video understanding ability of a test model. You are given the meta information about a video that includes:

2280  
2281  
2282  
2283  
2284  
2285

- Video captions: A complete description of the entire video for your reference.
- Scene descriptions: Detailed descriptions of key scenes throughout the video.
- Observed Scenes: Scenes in the given video that the test model can observe.
- Last Scene: The last scene of the entire video.

2286  
2287

Requirements:

2288

1. Question Content:

2289  
2290  
2291

- Given the video with observed scenes (scene 1 to k), the question should force the test model to predict future events (scene k+1 to scene n) and ask what intermediate events would be supposing scene n is given and scene n is the potential future end.

2292  
2293

- For example, "Question": "Based on the given video, predict future events and fill in the potential events in the given future events:  
1. [?] 2. [?] 3. [?] 4. [describe scene n]. "Options": A/B/C/D.  
[describe scene for slot 1], [describe scene for slot 2] [describe scene for slot 3]

2294  
2295  
2296

- Keep the event slot [?] to be filled.
- Construct the future event gap so that it is hard enough. For example, wrong answers could present the wrong order of the predicted future events.

2297  
2298  
2299

- Avoid using scene id in the question and start the question from "Based on the given video, ..."

2300

2. Question Format:

2301  
2302  
2303

- Create one multiple-choice question with four answer options: A, B, C, and D.
- Ensure only one correct answer and that the remaining three options are wrong.
- Only output required question-answer pairs shown in the output structure.

2304  
2305

Output structure:

2306  
2307

```
{output_structure}
```

2308  
2309

Please generate an example question based on the following input data.

2310  
2311

Input Data:

2312  
2313  
2314

- Video captions: {caption}
- Scene descriptions: {event}
- Observed Scenes: {obs}
- Last Scene: {last}

2315  
2316

2317

2318

2319

2320

2321

2322

**FutureBench Interpolation Question Construction Prompt**

2323

2324

2325

You are an expert in video understanding. Your task is to generate one multiple-choice question to assess the video understanding ability of a test model. You are given the meta information about a video that includes:

2326

- Video captions: A complete description of the entire video for your reference.
- Scene descriptions: Detailed descriptions of key scenes throughout the video.
- Observed Scenes: Scenes in the given video that the test model can observe.
- Last Scene: The last scene of the entire video.

2327

2328

2329

2330

2331

2332

2333

2334

2335

Requirements:

2336

1. Question Content:

2337

- Given the video with observed scenes (scene 1 to k), the question should force the test model to predict future events (scene k+1 to scene n) and ask what intermediate events would be supposing (scene k+i and scene k+j are given, k+i and k+j are potential future events).
- For example, "Question": "Based on the given video, predict future events and fill in the potential events in the given future events:  
1. [describe scene k+1] 2. [?] 3. [describe scene k+i] 4. [?] 5. [describe scene k+j]. "Options": A) [describe scene k+2], [describe scene k+j-1] B) [describe scene k+j-1], [describe scene k+2] C) [describe scene k+i+1], [describe scene k+i-1] D) [describe scene k+i-1], [describe scene k+2]
- Formulate the question so that the test model would not be able to deduce the correct answer without the observed scenes.
- Formulate the question so that it is hard enough and the test model would not be able to deduce the correct answer with only commonsense knowledge.
- Avoid using scene id in the question and start the question from "Based on the given video, ..."

2338

2339

2340

2341

2342

2343

2344

2345

2346

2347

2348

2349

2350

2351

2352

2. Question Format:

2353

- Create one multiple-choice question with four answer options: A, B, C, and D.
- Answer options should be built upon the scenes after the observed scenes and before the last scene.
- Ensure only one correct answer and that the remaining three options are wrong.
- Ensure each wrong answer contains related information to the observed scene but include missing details or only part of them are correct.
- Only output required question-answer pairs shown in the output structure.

2354

2355

2356

2357

2358

2359

2360

2361

2362

2363

2364

Output structure:

2365

{output\_structure}

2366

Please generate an example question based on the following input data.

2367

2368

Input Data:

2369

- Video captions: {caption}
- Scene descriptions: {event}
- Observed Scenes: {obs}
- Last Scene: {last}

2370

2371

2372

2373

2374

2375

## 2376 F IMPLEMENTATION DETAILS

2378 We conducted our experiments using two open-source frameworks: *LLaMA-Factory* [Zheng et al.](#) (2379 2024) for supervised video instruction tuning, and *EasyR1* [Zheng et al. \(2025\)](#) (based on the Verl 2380 framework [Sheng et al. \(2024\)](#)), optimized for reinforcement learning with multimodal data. 2381

2382 For supervised video instruction tuning, we trained our Qwen2.5-3B-VL-Instruct and Qwen2.5-7B- 2383 VL-Instruct models using LLaMA-Factory. Both models were fine-tuned for three epochs on 8 2384 NVIDIA A100 GPUs, employing the AdamW optimizer with a cosine learning rate scheduler, an 2385 initial learning rate of  $1 \times 10^{-5}$ , and a warm-up ratio of 0.1 to ensure stable training dynamics. To 2386 optimize memory usage and address computational constraints, each GPU processed one training 2387 sample per step, with gradient accumulation every two steps, effectively simulating a larger total 2388 batch size of 16 (2 steps  $\times$  8 GPUs). 2389

2390 For the reinforcement learning phase, experiments were executed using EasyR1 to further enhance 2391 model capabilities through multimodal refinement learning with Group Relative Policy Optimization 2392 (GRPO) [Guo et al. \(2025\)](#); [Shao et al. \(2024\)](#), also utilizing 8 NVIDIA A100 GPUs. We fine-tuned 2393 the Qwen2.5-VL-7B-Instruct model with a maximum prompt length of 4096 tokens and a response 2394 length capped at 2048 tokens. Training utilized global batch sizes of 16 samples per rollout, with 2395 micro-batches of four samples per GPU during parameter updates and eight per GPU for experience 2396 collection. We set the entropy coefficient to  $1 \times 10^{-3}$  to encourage exploration and the KL-divergence 2397 loss coefficient to  $1 \times 10^{-2}$  to maintain stable policy updates. Rollouts were configured to run eight 2398 steps without tensor parallelism or chunked prefill, ensuring efficient training and stable convergence. 2399 Evaluation logging was performed periodically, capturing ten generations per validation. Model 2400 checkpoints were systematically saved every 200 training iterations for comprehensive monitoring. 2401

2402 To accommodate varied visual inputs, in both settings, image resolutions were constrained between a 2403 minimum of 3136 pixels and a maximum of 1,605,632 pixels, ensuring consistency and computational 2404 manageability across diverse multimodal data. 2405

2406 For evaluation, we leveraged the open-source multimodal evaluation framework *lmms-eval* [Zhang et al. \(2024a\)](#). The framework encompasses all the benchmarks except SeedBench-R1 and our 2407 proposed FutureBench. To enhance reproducibility and usability, we integrated both SeedBench-R1 2408 and FutureBench into the *lmms-eval* framework. Hyperparameters followed the default settings 2409 provided by *lmms-eval*. 2410

## 2411 G LIMITATION AND FUTURE WORK

2412 Despite demonstrating the effectiveness of Next-Event Prediction (NEP) in advancing temporal 2413 reasoning capabilities in Multimodal Large Language Models (MLLMs), our current work has several 2414 limitations that invite further exploration. First, NEP primarily relies on automatically generated 2415 textual descriptions for future video segments as supervision signals. Although this approach offers 2416 scalability and avoids costly human annotations, the quality of generated captions might not match 2417 human-level precision and may reflect biases inherent in the annotation models used (e.g., GPT- 2418 4o [Hurst et al. \(2024\)](#)). Future research could explore integrating annotations from diverse sources, 2419 such as human annotators or alternative advanced models like Gemini [Team \(2025\)](#), to enhance 2420 annotation quality and reduce biases. 2421

2422 Second, while our proposed V1-33K dataset encompasses diverse scenarios, it may not fully capture 2423 all possible real-world video contexts, particularly highly specialized or infrequent event sequences. 2424 Extending this dataset by including additional domains, incorporating larger datasets, or employing 2425 synthetic video generation techniques could further enhance the diversity and robustness of the dataset, 2426 thereby strengthening models' temporal reasoning abilities. [In particular, our current NEP data are 2427 constructed from the \*intersection\* of video corpora that already support captioning and QA-style tasks 2428 in order to enable a fair task-level comparison. This design choice reduces the effective diversity 2429 of video sources and, together with the limited set of temporal benchmarks we evaluate on, may 2430 contribute to the observed saturation of downstream performance when scaling NEP data beyond 5K 2431 samples.](#) 2432

2433 Third, current state-of-the-art (SOTA) models often integrate diverse instruction-tuning datasets 2434 and tasks and leverage model merging strategies to optimize performance across benchmarks. [By](#) 2435

contrast, our experimental protocol primarily focuses on comparing different tasks individually and deliberately trains NEP in isolation, without combining it with QA or captioning data. Since most evaluation benchmarks adopt a QA-style format, this setup places NEP at a slight disadvantage and may underestimate its potential benefits in more realistic multi-task training recipes. Future research aimed at achieving SOTA performance across a wider array of benchmarks could benefit from (i) collecting more diverse video sources, (ii) scaling NEP to larger and less restricted corpora, and (iii) exploring combined instruction-tuning strategies and model merging approaches that jointly optimize NEP with QA and captioning objectives under a carefully designed curriculum.

Addressing these limitations will significantly enhance the reliability, generalizability, and depth of temporal reasoning capabilities in video-based multimodal language models.

## H BROADER IMPACTS

The proposed next-event prediction task has the potential to have a significant positive societal impact by improving multimodal models’ temporal reasoning capabilities, increasing their utility in applications such as video-based surveillance, assistive technology, and educational content generation. Improved predictive understanding of dynamic events could also help in safety-critical situations like traffic management and emergency response systems. However, there are some drawbacks, such as the risk of reinforcing biases embedded in training datasets, which is exacerbated by the reliance on automatically generated captions without human oversight. Careful consideration, transparent documentation, and strict ethical oversight will be essential to mitigate these risks and ensure responsible deployment.

## I LICENSES

We use standard licenses from the community. We include the following licenses for the codes, datasets and models we used in this paper.

Datasets & Benchmarks:

- VideoMME (Fu et al., 2024): [CC BY-NC 4.0](#)
- MVBench (Li et al., 2024d): [MIT](#)
- LongVideoBench (Wu et al., 2024): [CC-BY-NC-SA 4.0](#)
- TemporalBench (Cai et al., 2024): [MIT](#)
- TempCompass (Liu et al., 2024b): [CC BY-NC 4.0](#)
- SeedBench-R1 (Chen et al., 2025): [Apache License 2.0](#)
- LLaVA-Video-178K (Zhang et al., 2023): [Apache License 2.0](#)

Codes:

- verl (Sheng et al., 2024): [Apache License 2.0](#)
- EasyR1 (Zheng et al., 2025): [Apache License 2.0](#)
- LLaMA-Factory Zheng et al. (2024): [Apache License 2.0](#)

Models:

- Qwen2.5-VL-7B-Instruct (Bai et al., 2025): [Apache License 2.0](#)
- Qwen2.5-VL-3B-Instruct (Bai et al., 2025): [Apache License 2.0](#)
- OpenAI API (Hurst et al., 2024): [OpenAI API Terms of Use](#)

## J LLM USAGE

We used an OpenAI LLM (GPT-5) as a writing and formatting assistant. In particular, it helped refine grammar and phrasing, improve clarity, and suggest edits to figure/table captions and layout (e.g.,

2484 column alignment, caption length, placement). The LLM did not contribute to research ideation, exper-  
2485 imental design, implementation, data analysis, or technical content beyond surface-level edits. All out-  
2486 puts were reviewed and edited by the authors, who take full responsibility for the final text and visuals.  
2487

2488

2489

2490

2491

2492

2493

2494

2495

2496

2497

2498

2499

2500

2501

2502

2503

2504

2505

2506

2507

2508

2509

2510

2511

2512

2513

2514

2515

2516

2517

2518

2519

2520

2521

2522

2523

2524

2525

2526

2527

2528

2529

2530

2531

2532

2533

2534

2535

2536

2537