

Imagine Before You Predict: Interleaved Latent Visual Reasoning for Video Event Prediction

Tianxiang Jiang^{*1,2} Linquan Wu^{*3} Sheng Xia⁴ Songze Li^{5,2}
Ziang Yan^{6,2} Haoyu Yang⁷ Yu Qiao² Yi Wang^{2†}

¹ University of Science and Technology of China ² Shanghai AI Laboratory
³ City University of Hong Kong ⁴ Nanjing University ⁵ Fudan University
⁶ Zhejiang University ⁷ University of Electronic Science and Technology of China
<https://github.com/OpenGVLab/Future-L1>

Abstract

Video event prediction (VEP) requires models to infer unobserved future states from partial video evidence. Existing video MLLMs usually verbalize intermediate future reasoning in text space: once visual evidence is verbalized, fine-grained motion, geometry, and interaction cues can be lost, leading to plausible but visually ungrounded hallucinations. We introduce **FUTURE-L1**, an interleaved latent visual reasoning framework that lets an MLLM alternate between language tokens and continuous latent visual spans during autoregressive decoding. To train this capability, we construct **FUTURE-L1-50K** by selecting examples where future visual hints help prediction and align latent states to future-frame embeddings, then further optimize sampled latent trajectories with **LA-DAPO**, a latent-aware RL objective with outcome-contrastive and temporal-diversity rewards. **FUTURE-L1** achieves new *state-of-the-art* results on both benchmarks: on FutureBench, it improves Qwen3-VL-8B from 61.0 to **85.4** and exceeds the previous best Video-CoE by 10.4 points; on TwiFF-Bench, it improves the average score from 2.44 to **3.04**. These results suggest that future-oriented video reasoning benefits from preserving intermediate visual semantics in latent space rather than translating every reasoning step into text.

1 Introduction

Video event prediction (VEP) asks a model to infer what will happen next from a partially observed video (Koppula and Saxena, 2016; Vondrick et al., 2016a; Lei et al., 2020; Wang et al., 2025b; Su et al., 2026). Unlike standard video understanding, whose answers can usually be grounded in visible frames, VEP requires constructing an internal hypothesis about unobserved dynamic visual states:

where objects will move, whether entities will interact, and how a scene will evolve. Although recent multimodal large language models (MLLMs) have made rapid progress on retrospective video tasks (Bai et al., 2025b,a; Wang et al., 2024; Li et al., 2024; Fu et al., 2025; Li et al., 2025c), future-oriented reasoning remains less explored.

Existing video MLLMs usually verbalize intermediate future reasoning in text space (Zhang et al., 2023; Han et al., 2025; Feng et al., 2025; Li et al., 2025d; Su et al., 2026). This is convenient for explanation, but it creates a poor interface for dynamic visual prediction: once visual evidence is converted into words, fine-grained motion, geometry, relative position, and interaction can be lost. The resulting reasoning may sound plausible while drifting away from visual semantics, especially when the correct answer depends on subtle future dynamics. Recent latent visual reasoning methods avoid part of this bottleneck by using continuous visual states (Li et al., 2025b; Pham and Ngo, 2025; Qin et al., 2025; Cheng et al., 2026; Li et al., 2025a; Yang et al., 2025b; Lu et al., 2026a), but most treat latent thoughts as static helper images or one-shot visual hints. VEP instead calls for a temporally organized latent process that can update imagined dynamic visual states over multiple reasoning steps.

We introduce **FUTURE-L1**, a framework that equips MLLMs with **interleaved latent visual reasoning** for VEP. During autoregressive decoding, **FUTURE-L1** alternates between textual tokens and continuous latent visual spans, allowing language to organize the reasoning while latent states preserve intermediate dynamic visual structure. Training proceeds in two stages. First, we construct **FUTURE-L1-50K** from TwiFF-style trajectories using visual-gain data curation, selecting examples where intermediate future visual hints measurably help prediction. Supervised fine-tuning then teaches the model when to invoke latent spans and aligns their hidden states with future-frame em-

† Corresponding author. This work was performed at Shanghai AI Laboratory. * Equal contribution.

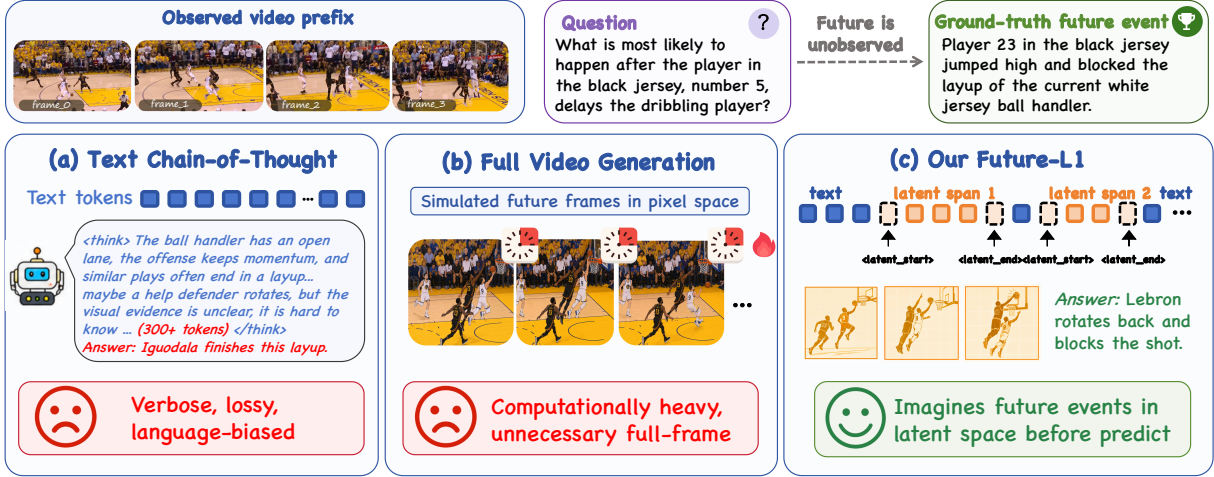


Figure 1: **Motivation of interleaved latent visual reasoning.** Text-CoT can be verbose and visually lossy, while pixel-space future simulation is computationally heavy. **FUTURE-L1** instead inserts compact latent visual spans that preserve dynamic future semantics without generating full frames.

beddings. Second, we apply **LA-DAPO**, a latent-aware RL objective that optimizes sampled latent trajectories with outcome-contrastive and temporal-diversity rewards, encouraging successful latent futures while discouraging repeated visual thoughts.

Experiments show that latent visual reasoning is substantially more effective than text-only reasoning for VEP. On FutureBench, FUTURE-L1-RL improves Qwen3-VL-8B from 61.0 to **85.4**, exceeding the previous best Video-CoE by **10.4** points. On TwiFF-Bench, it improves the average score from 2.44 to **3.04**. Under the same curated data source, text-only SFT reaches only 65.0 on FutureBench, whereas interleaved latent SFT reaches **73.2**, indicating that the gain is not merely from additional supervision but from reasoning through a modality better matched to future visual structure.

Our contributions are threefold:

1. We propose visual-gain data curation and construct **FUTURE-L1-50K**, a high-utility corpus for supervising latent future visual reasoning.
2. We introduce **interleaved latent visual reasoning for VEP**, enabling autoregressive models to alternate between language and continuous future visual states.
3. We develop **LA-DAPO**, a latent-aware RL method that improves sampled latent trajectories and achieves **state-of-the-art** results on FutureBench and TwiFF-Bench.

2 Related Work

Multimodal Large Language Models. Multimodal large language models (MLLMs) connect

visual encoders with strong LLM backbones and have become the mainstream framework for visual understanding (Bai et al., 2025a; Team et al., 2026a; Hong et al., 2026; Xiao et al., 2026; An et al., 2026). For video understanding, recent MLLMs extend image-based models with temporal frame sampling, video instruction tuning, longer-context modeling, and large-scale video-text corpora (Wang et al., 2024; Zhang et al., 2024; Wang et al., 2025a; Yan et al., 2026), substantially improving performance on diverse benchmarks (Li et al., 2024; Fu et al., 2026; Yang et al., 2025a; Xu et al., 2025; Shi et al., 2026; Liu et al., 2026b). Beyond perception and recognition, reasoning-oriented post-training has been applied to MLLMs, including chain-of-thought supervision (Han et al., 2025) and reinforcement learning (Li et al., 2025d). More recently, paradigms that encourage models to *think with images or videos* move beyond purely textual rationales by retrieving visual evidence (Zheng et al., 2025b; Zeng et al., 2026; Wang et al., 2026; Lu et al., 2026b; Wu et al., 2026b) with visual traces, motivating non-textual intermediate representations for visual reasoning.

Reasoning in Latent Space. Latent reasoning (Yu et al., 2026) replaces discrete textual reasoning tokens with continuous hidden states fed back into the LLM, compressing chain-of-thought into a compact thinking space. Coconut (Hao et al., 2024) first showed that an LLM can reason in its own embedding space, and CODI (Shen et al., 2025) and SIM-CoT (Wei et al., 2025) subsequently distilled or supervised these latent steps to

close the gap to explicit textual CoT. This paradigm has also been adopted by MLLMs through visual supervision: Mirage (Yang et al., 2025b) and LVR (Li et al., 2025a) align latent slots with embeddings of helper images that hint at the answer, and LaViT (Wu et al., 2026a) further constrains latent visual thoughts with teacher-guided attention. More flexible designs allow models to alternate between textual tokens and continuous visual states during reasoning, as in Monet (Wang et al., 2025d), SkiLa (Tong et al., 2025), and SwimBird (Tong et al., 2026). However, these methods largely anchor latent thoughts to *static images*, such as helper images, sketches, or scenes already given to the model. Video event prediction instead requires reasoning over *dynamic future frames* that are not yet observed, where above studies have not explored. FUTURE-L1 accordingly grounds latent thoughts in future information rather than static visual hints.

Video Event Prediction. Unlike standard video understanding benchmarks (Li et al., 2024; Fu et al., 2025; Liu et al., 2024) that focus on visible content, video event prediction requires models to infer unobserved future events from a video prefix. This future-oriented setting spans low-level action anticipation (Lan et al., 2014; Gammulle et al., 2019), future-frame prediction (Ranzato et al., 2014; Vondrick et al., 2016b), and high-level semantic next-event prediction (Lei et al., 2020; Jiang et al., 2025; Liang et al., 2025; Su et al., 2025). Most VEP methods remain text-output oriented (Cheng et al., 2025a; Wang et al., 2025b); for example, VideoCoE (Su et al., 2026) structures the reasoning trace as a long textual chain of historical events. Videoas-Answer (Cheng et al., 2025b) instead moves the answer modality from text to generated video explicitly. FUTURE-L1 differs from these routes: rather than verbalizing every intermediate event or synthesizing full videos, it represents intermediate future states in an interleaved latent visual channel supervised by future-frame embeddings.

3 Method

We propose FUTURE-L1, an interleaved latent visual reasoning framework for VEP. Given an observed video prefix V and question q , the model generates a response y by alternating textual reasoning, bounded latent visual spans, and a final answer. Training has two stages: SFT on FUTURE-L1-50K teaches when to invoke latent spans and aligns them with future-frame embeddings, while

LA-DAPO further optimizes sampled latent trajectories with outcome-contrastive and temporal-diversity rewards. Figure 2 illustrates the pipeline.

3.1 Interleaved Latent Visual Reasoning

Autoregressive Reasoning with Latent Visual Spans. FUTURE-L1 augments a standard MLLM backbone (Bai et al., 2025a) with a latent visual reasoning channel using three special tokens: `<|latent_start|>`, `<|latent|>`, and `<|latent_end|>`. Generation begins in textual mode. Once `<|latent_start|>` is emitted, each following `<|latent|>` position produces a hidden state h_t that is fed back as the next input embedding rather than projected to the vocabulary. These continuous states act as latent visual thoughts and remain in the KV cache to condition later textual reasoning. Generation returns to text when `<|latent_end|>` is emitted.

Dynamic Latent Budget at Inference. Latent span length is not fixed: a span ends when the model emits `<|latent_end|>`. We cap each span by $L_{\max}$ to avoid run-on latent decoding, and a response may contain multiple spans, allowing the model to allocate latent computation adaptively across reasoning stages.

3.2 SFT with FUTURE-L1-50K

SFT provides a necessary cold start for latent reasoning by training on curated interleaved traces and aligning latent states with future-frame embeddings. This prevents the model from either avoiding latent spans or producing continuous states not grounded in meaningful visual manifold before RL.

Visual-Gain Data Curation. We curate **FUTURE-L1-50K** from TwiFF-2.7M (Liu et al., 2026a), a VCoT corpus that provides intermediate reasoning frames. Unlike synthesized sketches or generic helper images, these frames are temporally later frames from the same authentic video, so they depict unseen future states that are physically consistent with the observed prefix. This makes them a natural supervision signal for latent visual reasoning: the model is not asked to imitate arbitrary visual hints, but to internalize future visual states that actually occur.

However, not every TwiFF sample provides useful supervision for VEP. Some examples are already easy to solve from the observed prefix alone, where extra future-frame hints add little value. Others remain ambiguous or uninformative even when

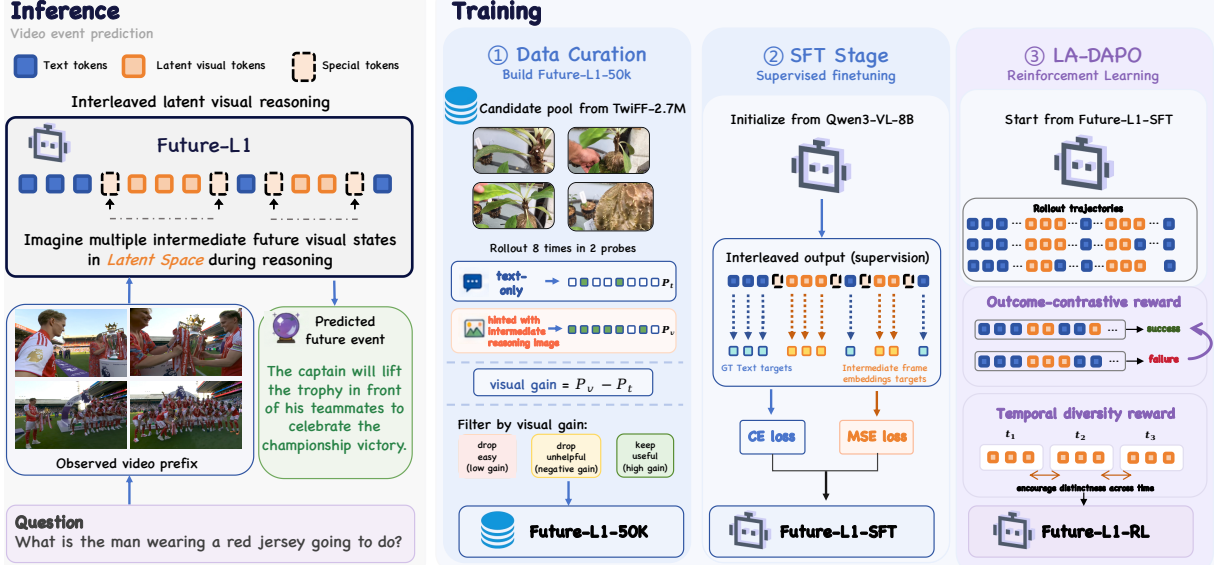


Figure 2: **Overview of FUTURE-L1.** (Left) FUTURE-L1-50K is built by ranking TwiFF candidates by visual gain $p_v - p_t$. (Center) SFT trains interleaved text-latent trajectories, aligning latent spans with future visual states. (Right) LA-DAPO further optimizes sampled trajectories with outcome-contrastive and temporal-diversity rewards.

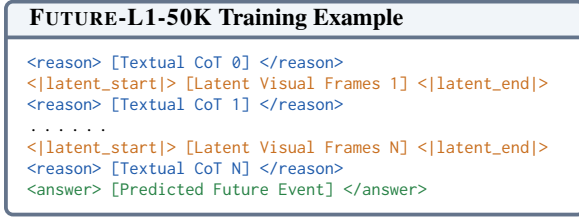


Figure 3: FUTURE-L1-50K training format: textual reasoning interleaved with bounded latent visual spans supervised by future-frame embeddings.

a reasoning frame is provided. Training on them dilutes the signal that latent visual states should carry. We therefore filter examples by the *marginal utility* of their intermediate reasoning frames.

For each candidate, we evaluate Qwen3-VL-8B-Instruct under two conditions: (1) a text-only input with the observed video prefix and question; and (2) a hinted input that additionally includes the intermediate reasoning frames. Each condition uses 8 independent rollouts judged by Qwen3.5-397B-A17B. Let $p_t, p_v \in [0, 8]$ be the correct-rollout counts; we **retain** samples with $p_t \leq 6$, so the text-only setting is not saturated, and $p_v - p_t \geq 2$, so the visual hint provides measurable lift. We **rank** retained samples by descending $p_v - p_t$, and take the top 50,000 items as **FUTURE-L1-50K**. All retained samples are reformatted into the interleaved trajectory shown in Figure 3.

Training Objective. SFT optimizes a joint objective over discrete text tokens and continuous latent

visual states:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{Latent}}, \quad (1)$$

where λ controls the strength of latent supervision.

For discrete positions $\mathcal{T}$, including textual reasoning, answer tokens, and special control tokens, we use standard next-token prediction:

$$\mathcal{L}_{\text{CE}} = - \sum_{t \in \mathcal{T}} \log p_{\theta}(w_t \mid w_{<t}, V, q). \quad (2)$$

For latent positions $\mathcal{S}$, we align each hidden state $\mathbf{h}_t$ with the visual embedding $\mathbf{e}_t^*$ of the corresponding future reasoning frame, extracted by the Qwen3-VL vision encoder:

$$\mathcal{L}_{\text{Latent}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \|\mathbf{h}_t - \mathbf{e}_t^*\|_2^2. \quad (3)$$

This anchors latent spans to the future-frame manifold while preserving standard language modeling over the textual channel.

3.3 LA-DAPO for Latent-Aware RL

SFT provides a grounded but teacher-forced initialization: each latent state is matched to a future-frame embedding, while sampled latent trajectories are not directly optimized for prediction success. We therefore introduce **LA-DAPO** (Latent-Aware Direct Advantage Policy Optimization), a latent-aware extension of DAPO (Yu et al., 2025). LA-DAPO keeps DAPO’s answer and format rewards, and adds two trajectory-level latent rewards:

an **outcome-contrastive reward** that aligns latent trajectories associated with correct answers, and a **temporal-diversity reward** that discourages repeating the same visual thought across spans. Because these rewards depend on rollout outcomes and generated latent states, LA-DAPO can optimize latent reasoning **without requiring intermediate-frame annotations** during RL.

Outcome-Contrastive Latent Reward. Answer rewards provide only a sequence-level scalar, leaving latent states weakly supervised. We introduce an **outcome-contrastive reward** R_{ctr} that structures latent trajectories by group outcomes: correct rollouts are pulled together, while incorrect rollouts serve as negatives. Because the signal depends only on final-answer correctness, it does not require intermediate-frame annotations.

Let $\mathbf{Z}_i = [\mathbf{z}_{i,1}, \dots, \mathbf{z}_{i,T_i}]$ be the normalized latent trajectory of rollout i , with correctness $a_i \in \{0, 1\}$. We define trajectory similarity as

$$s_{ij} = \frac{1}{T} \sum_{t=1}^T \frac{1 + \langle \mathbf{z}_{i,t}, \mathbf{z}_{j,t} \rangle}{2}, \quad (4)$$

where $T = \min(T_i, T_j)$. Let $\mathcal{P}_i = \{j \neq i : a_j = 1\}$, $\mathcal{N}_i = \{j \neq i : a_j = 0\}$, and $s_i^+ = \max_{j \in \mathcal{P}_i} s_{ij}$. We use a hardest-positive InfoNCE reward:

$$R_{\text{ctr}}(i) = \frac{\exp(s_i^+/\tau)}{\exp(s_i^+/\tau) + \sum_{j \in \mathcal{N}_i} \exp(s_{ij}/\tau)}. \quad (5)$$

Empty-positive or empty-negative cases are handled in Appendix B.

Temporal Diversity Reward. R_{ctr} aligns trajectories *across* rollouts but imposes no structure *within* a rollout: a policy can still earn a high answer reward by emitting near-identical latent states at consecutive spans, collapsing the latent channel into a single visual thought repeated over time. Although SFT discourages this through frame-distinct supervision, this constraint is no longer present during RL. We therefore add a **temporal diversity reward** R_{div} that encourages adjacent latent spans to represent distinct future updates. For a response with M latent spans, we mean-pool the latent vectors within span m into a representative $\mathbf{b}_m$, and penalize adjacent-span similarity:

$$R_{\text{div}} = -\frac{1}{M-1} \sum_{m=1}^{M-1} \cos^2(\mathbf{b}_m, \mathbf{b}_{m+1}). \quad (6)$$

This reward is maximized at 0 when adjacent span representatives are orthogonal and decreases as they become redundant.

Together, R_{ctr} and R_{div} regularize latent reasoning along complementary axes: R_{ctr} links latent trajectories to prediction outcomes across rollouts, while R_{div} keeps successive latent spans within a rollout temporally distinct.

Final Rewards. The total target combines answer / format rewards and two latent terms,

$$R = \lambda_a R_{\text{acc}} + \lambda_f R_{\text{fmt}} + \lambda_c R_{\text{ctr}} + \lambda_d R_{\text{div}}, \quad (7)$$

where λ_a and λ_f follow the DAPO defaults (Appendix B); λ_c and λ_d are ablated in §4.

4 Experiments

Benchmarks. We evaluate FUTURE-L1 on two complementary video event prediction benchmarks. *FutureBench* (Wang et al., 2025b) is a multiple-choice VEP benchmark that asks models to predict unobserved future events from a video prefix. It reports overall accuracy and four reasoning-depth splits: *1-Hop*, *2-Hop*, *3-Hop*, and *Interp.*. While *1-Hop* mainly tests immediate next-event prediction, *3-Hop* and *Interp.* form harder OOD-style regimes: *3-Hop* requires extrapolating longer future event chains, and *Interp.* requires reasoning over non-consecutive future states under partial intermediate anchors. These splits therefore test whether a model can generalize beyond local next-event cues. *Twiff-Bench* (Liu et al., 2026a) evaluates open-ended future-frame reasoning over 1,078 QA samples and scores both the generated reasoning trajectory and the final answer. Following the official protocol, we report CoT quality, answer quality, and their average under the benchmark judge. The Twiff-Bench evaluation set is not used in FUTURE-L1-50K construction, SFT, or RL training.

Implementation Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025a) as the backbone. SFT trains for 1 epoch on FUTURE-L1-50K (§3.2) with global batch size 128, peak learning rate 1×10^{-5} , MSE weight $\lambda=0.1$, and maximum latent budget $L_{\text{max}}=4$ unless otherwise specified. RL starts from the SFT checkpoint with group size $G=8$ and uses Qwen3.6-27B as the LLM-as-judge for the accuracy reward. All experiments run on $8 \times \text{NVIDIA H200 GPUs}$, and all checkpoints are evaluated with *lmms-eval* (Zhang et al., 2025a). More detailed settings are listed in Appendix C.

Table 1: **Main results on FutureBench** (Wang et al., 2025b). Accuracy (%); best results are in **bold**.

Model	Size	Method	Frames	1-Hop	2-Hop	3-Hop	Interp.	AVG
<i>Open-source and Proprietary Models</i>								
GLM-4.1V (Team et al., 2026b)	9B		32	29.9	41.9	52.2	47.3	44.4
LLaVA-NeXT-Video (Zhang et al., 2025b)	7B		32	48.8	49.3	40.0	44.4	45.2
MiMo-VL (Xiaomi, 2025)	7B		32	59.0	59.6	50.5	43.8	50.5
InternVL3 (Zhu et al., 2025a)	8B		32	54.3	58.0	63.2	54.4	56.7
Qwen2.5-VL-Instruct (Bai et al., 2025b)	7B	Zero-Shot	32	57.2	57.0	50.2	50.7	52.9
Qwen2.5-VL-Instruct (Bai et al., 2025b)	72B		32	55.5	68.4	63.7	53.2	58.3
Qwen3-VL (Bai et al., 2025a)	30B-A3B		32	65.3	70.5	76.1	62.2	66.9
GPT-4o (OpenAI, 2024)	–		32	61.9	61.7	72.1	51.6	59.0
GPT-5 (Singh et al., 2026)	–		32	59.6	57.3	62.6	55.6	57.9
<i>Video Reasoning Models</i>								
Video-RFT (Wang et al., 2025c)	7B	SFT+RL	32	62.4	53.9	50.7	53.8	54.6
Video-R1 (Feng et al., 2025)	7B	SFT+RL	32	67.6	65.3	61.2	61.8	63.3
VideoAuto-R1 (Liu et al., 2026c)	8B	SFT+RL	32	63.6	69.4	67.7	59.3	63.4
Video-o3 (Zeng et al., 2026)	7B	SFT+RL	32	68.2	73.6	63.2	69.7	68.9
NEP (Wang et al., 2025b)	7B	SFT+RL	32	66.2	69.9	63.7	68.1	67.3
Video-CoE (Su et al., 2026)	7B	SFT+RL	32	80.9	83.9	71.6	71.4	75.0
<i>Latent Visual Reasoning Models</i>								
LVR (Li et al., 2025a)	7B	SFT+RL	32	22.5	26.4	22.9	17.6	21.0 [†]
Monet (Wang et al., 2025d)	7B	SFT+RL	32	46.8	47.2	45.3	49.7	47.9
SwimBird (Tong et al., 2026)	8B	SFT	32	59.0	66.8	64.7	61.8	62.8
<i>Ours</i>								
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	Zero-Shot	32	64.2	65.8	66.2	55.8	61.0
Text-Only SFT (on FUTURE-L1-50K)	8B	SFT	32	67.6	66.8	68.2	62.0	65.0
FUTURE-L1	8B	SFT	32	70.5	73.1	77.6	72.2	73.2
FUTURE-L1	8B	SFT+RL	32	83.2	86.5	86.6	85.1	85.4

[†]LVR often collapses under dense video visual-token inputs and fails to produce valid text responses.

Table 2: **Main results on TwiFF-Bench** (Liu et al., 2026a). Avg.=(CoT+Ans)/2; best results are in **bold**.

Model	Size	CoT	Answer	Avg.
<i>Multimodal Large Language Models</i>				
Qwen2.5-VL (Bai et al., 2025b)	7B	2.46	1.63	2.05
InternVL3.5 (Wang et al., 2025e)	8B	2.35	1.85	2.10
DeepEyes (Zheng et al., 2025b)	7B	2.54	2.20	2.37
<i>Unified Models</i>				
Janus-Pro (Chen et al., 2025)	7B	2.04	1.04	1.54
Bagel (Deng et al., 2025)	7B	2.29	1.85	2.07
TwiFF-300K (Liu et al., 2026a)	7B	2.90	2.55	2.73
TwiFF-2.7M (Liu et al., 2026a)	7B	2.95	2.62	2.79
<i>Ours</i>				
Zero-Shot (Bai et al., 2025a)	8B	2.75	2.14	2.44
FUTURE-L1-SFT	8B	2.62	2.42	2.52
FUTURE-L1-RL	8B	3.11	2.97	3.04

4.1 Main Results

Prior Models Struggle on VEP. Tables 1 and 2 show that VEP remains difficult even for strong MLLMs. Proprietary and open-source models do not reliably solve FutureBench: GPT-4o obtains 59.0, GPT-5 obtains 57.9, and Qwen3-VL-30B-A3B reaches 66.9. Video-reasoning models improve over generic MLLMs but continue to struggle,

including Video-R1 (63.3), Video-o3 (68.9), NEP (67.3), and Video-CoE (75.0). Their remaining errors are especially visible on the harder future-oriented splits: the strongest Video-CoE reaches only 71.6 on 3-Hop and 71.4 on Interp., where models must extrapolate **longer** event chains or reason over **non-consecutive** future states. Existing **static** latent visual reasoning methods also do not transfer directly to dense video prediction: Monet reaches 47.9 and LVR obtains 21.0. These results suggest that VEP is not solved by scaling generic MLLMs, adding text-centric video reasoning, or directly reusing static latent-reasoning recipes.

FUTURE-L1 Boosts FutureBench. FUTURE-L1-SFT reaches **73.2**, improving the Qwen3-VL backbone (from 61.0) by **+12.2**. It outperforms the text-only SFT control trained on the same FUTURE-L1-50K (65.0) by **8.2**, isolating the gain from interleaved latent reasoning rather than sample selection alone. After LA-DAPO, FUTURE-L1-RL improves to **85.4**, exceeding Qwen3-VL-30B-A3B by **18.5** points and Video-CoE by **10.4** points. The gains over the backbone are strongest on the harder splits: **+19.0**, **+20.7**, **+20.4**, and **+29.3** on 1-Hop,

Table 3: **SFT hyperparameter ablation on FutureBench.** Accuracy (%) for latent MSE weight λ and budget $L_{\max}$.

Setting	1-Hop	2-Hop	3-Hop	Interp.	AVG
<i>Latent MSE weight λ</i>					
0.01	68.2	69.9	73.1	67.5	69.1
0.05	71.1	72.0	73.6	69.3	70.9
0.10	70.5	73.1	77.6	72.2	73.2
0.20	69.9	76.7	74.6	70.1	72.2
0.50	73.4	71.0	71.6	69.3	70.7
1.00	73.4	73.1	68.7	67.1	69.5
<i>Maximum latent budget $L_{\max}$</i>					
2	66.5	74.1	74.6	69.3	70.7
4	70.5	73.1	77.6	72.2	73.2
8	65.9	75.1	73.6	72.4	72.1
16	69.9	72.5	71.1	70.8	71.0
32	69.4	72.0	71.1	69.5	70.3
64	67.1	68.9	70.6	65.6	67.4

2-Hop, 3-Hop, and Interp., respectively. The larger improvements on 3-Hop and Interp. suggest that latent channel generalizes to longer future chains, rather than only improving single-step NEP.

Twiff-Bench Shows the Same Trend. On Twiff-Bench, FUTURE-L1-SFT raises the average score from 2.44 to 2.52. Though its CoT score decreases from 2.75 to 2.62, its answer score rises from 2.14 to **2.42**, showing the curated traces strengthen prediction even when their surface reasoning is imperfect. LA-DAPO improves both dimensions, reaching **3.11** CoT and **2.97** Ans for an average of **3.04**. This surpasses the previous best Twiff-2.7M (2.79) and all listed MLLM or unified baselines, indicating that interleaved latent reasoning and trajectory-level RL are complementary.

4.2 Ablation Study

SFT Hyperparameters. Table 3 sweeps the latent MSE weight λ and the maximum latent budget $L_{\max}$. With $L_{\max} = 4$ fixed, $\lambda = 0.1$ is optimal (**73.2**); both weaker ($\lambda = 0.01$, 69.1) and stronger ($\lambda = 1.0$, 69.5) alignment weights cost 3-4 points, indicating that latent positions need explicit but not dominant supervision. With $\lambda = 0.1$ fixed, accuracy peaks at $L_{\max} = 4$ and degrades to **67.4** at $L_{\max} = 64$, suggesting that an overly long latent span dilutes useful signal. This indicates that latent reasoning benefits from short, explicitly supervised spans rather than simply allocating more continuous tokens. We adopt $\lambda = 0.1$, $L_{\max} = 4$ as the default SFT setting.

RL Objective. Table 4 ablates the RL objective from FUTURE-L1-SFT. GRPO (82.8) and

Table 4: **RL objective ablation on FutureBench.** Accuracy (%); all variants start from FUTURE-L1-SFT.

Method	1-Hop	2-Hop	3-Hop	Interp.	AVG
Text-Only SFT	67.6	66.8	68.2	62.0	65.0
+ GRPO	77.5	78.8	78.1	77.1	77.7
+ DAPO	83.2	81.3	78.1	71.2	76.3
FUTURE-L1-SFT	70.5	73.1	77.6	72.2	73.2
+ GRPO	82.7	84.5	85.1	81.2	82.8
+ DePO	78.0	80.3	86.6	80.2	81.1
+ DAPO	83.2	85.5	86.6	82.4	83.8
+ R_{ctr}	83.2	86.0	87.1	83.2	84.5
+ R_{div}	82.7	87.0	87.6	83.4	84.8
FUTURE-L1-RL	83.2	86.5	86.6	85.1	85.4

Table 5: **LA-DAPO reward coefficient ablation on FutureBench.** Accuracy (%) for λ_c and λ_d .

Setting	1-Hop	2-Hop	3-Hop	Interp.	AVG
<i>Outcome-contrastive weight λ_c</i>					
0.01	81.5	84.5	86.1	83.4	83.8
0.05	82.7	87.0	86.1	83.0	84.3
0.10	84.4	86.5	87.1	84.0	85.1
0.20	83.2	86.5	86.6	85.1	85.4
0.50	82.1	86.0	86.1	83.0	84.0
1.00	83.8	86.5	86.6	84.5	85.1
<i>Temporal diversity weight λ_d</i>					
0.01	83.2	86.5	86.6	83.8	84.8
0.05	83.8	87.0	86.6	84.3	85.1
0.10	83.2	86.5	86.6	85.1	85.4
0.20	80.9	82.9	87.1	83.2	83.5
0.50	79.8	83.4	85.6	81.6	82.4
1.00	78.0	82.4	85.6	81.0	81.6

DePO (Cheng et al., 2026) (81.1) already lift FUTURE-L1-SFT (73.2) by about 9 points, and DAPO further reaches **83.8**. Adding latent-aware rewards improves the objective beyond DAPO: the outcome-contrastive reward R_{ctr} raises performance to 84.5, the temporal-diversity reward R_{div} reaches 84.8, and using both in FUTURE-L1-RL achieves **85.4**. This shows that the gain is not only from stronger RL, but from rewards that directly structure latent visual trajectories.

RL Reward Coefficients. Table 5 examines the latent-reward coefficients. The outcome-contrastive weight peaks at $\lambda_c = 0.2$ (**85.4**), and the temporal-diversity weight peaks at $\lambda_d = 0.1$; larger values hurt, dropping to **81.6** at $\lambda_d = 1.0$. This suggests that contrastive alignment and temporal diversity are both useful, but excessive pressure can push latent spans off the manifold.

4.3 Analysis of Latent Visual Reasoning

Visual-Gain Filtering. Table 6 controls for a key confound: whether the SFT gain comes from visual-gain selection or merely from Twiff-style

Table 6: **Effect of visual-gain filtering.** FutureBench accuracy (%) for 50K TwiFF-format SFT data.

Training Set	1-Hop	2-Hop	3-Hop	Interp.	AVG
Zero-Shot	64.2	65.8	66.2	55.8	61.0
Random 50K	67.6	68.9	70.1	67.7	68.4
FUTURE-L1-50K	70.5	73.1	77.6	72.2	73.2

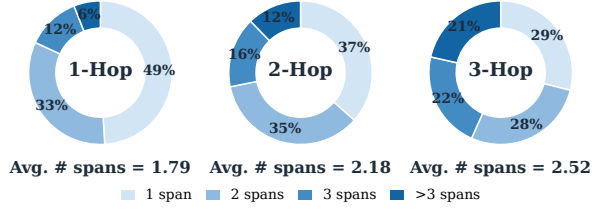


Figure 4: **Latent-span usage by reasoning depth.** Donuts show span-count distributions; values report mean spans over six RL settings.

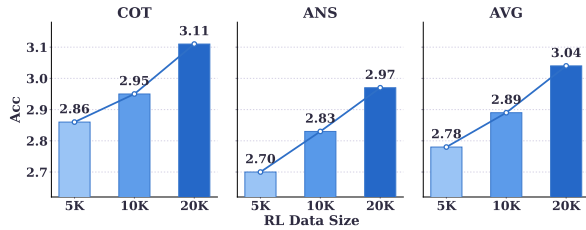


Figure 5: **RL data scaling on TwiFF-Bench.** Scores improve as LA-DAPO uses 5K, 10K, and 20K retained visual-gain samples.

formatting. We compare our Top-50K set with a random 50K set sampled from TwiFF-2.7M under the same interleaved-format requirement and train both with the same FUTURE-L1-SFT recipe. The random set improves Qwen3-VL-8B from 61.0 to 68.4, showing that interleaved demonstrations help, but it remains 4.8 points below our visual-gain selected set (73.2). The gap persists on the harder splits, including 3-Hop (70.1 vs. 77.6) and Interp. (67.7 vs. 72.2). Thus FUTURE-L1-50K improves transfer not only by exposing the model to TwiFF-style traces, but by selecting examples whose future visual hints provide measurable predictive utility.

Adaptive Latent Usage. Figure 4 examines whether FUTURE-L1 allocates latent computation according to reasoning difficulty (Zhu et al., 2025b). Averaged over six RL hyperparameter settings, the mean span count increases with depth, from 1.79 on 1-Hop to 2.18 on 2-Hop and 2.52 on 3-Hop. The distribution shifts in the same direction: one-span responses become less frequent as depth increases, while responses with more than three spans grow from 6% on 1-Hop to 12% on 2-Hop and 21% on 3-Hop. This shows that latent spans are not emitted as a fixed template; instead, FUTURE-L1 spends

Table 7: **Inference cost on FutureBench.** Average tokens, accuracy, latency, and accuracy per second.

Model	Tokens ↓	Acc. ↑	Latency (s) ↓	Acc./s ↑
Video-R1	398.5	63.3	3.28	19.3
Video-o3	348.6	68.9	25.90	2.7
Qwen3-VL-8B	288.8	61.0	1.18	51.7
FUTURE-L1-SFT	205.3	73.2	0.96	76.1
FUTURE-L1-RL	195.3	85.4	0.91	93.8

more latent visual budgets when longer future event chains require updating dynamic visual states.

RL Data Scaling. Figure 5 tests whether LA-DAPO benefits from more retained visual-gain data. Using 5K, 10K, and 20K samples from the retained pool, the TwiFF-Bench average score increases monotonically from 2.78 to 2.89 and 3.04. This trend indicates that trajectory-level latent RL continues to benefit from high-utility samples rather than saturating on a small preference set.

Inference Efficiency. Table 7 compares inference cost on FutureBench. Text-heavy and multi-turn baselines require substantially larger decoding budgets: Video-R1 emits 398.5 tokens at 3.28 seconds per sample, and Video-o3 emits 348.6 tokens at 25.90 seconds due to repeated model calls during search. In contrast, FUTURE-L1-SFT uses 205.3 tokens and reaches 73.1 accuracy at 0.96 seconds, while FUTURE-L1-RL uses 195.3 tokens and reaches 85.4 accuracy at 0.91 seconds, yielding the best accuracy-per-second score. Thus FUTURE-L1 improves accuracy through compact latent visual computation rather than expensive explicit multi-turn reasoning.

More analysis including latent visualizations and reward dynamics are provided in Appendix F.

5 Conclusion

We presented **FUTURE-L1**, an interleaved latent visual reasoning framework for video event prediction. The central idea is to keep dynamic future visual structure in a continuous latent channel instead of verbalizing every intermediate hypothesis as text. To make this practical, FUTURE-L1 first uses **FUTURE-L1-50K** to ground latent spans with future-frame embeddings selected by visual-gain curation, and then applies **LA-DAPO** to optimize sampled latent trajectories through outcome-contrastive and temporal-diversity rewards. Across FutureBench and TwiFF-Bench, this combination improves both multiple-choice future prediction and open-ended future reasoning, with especially

large gains on longer and non-consecutive future-event splits. These results suggest a broader direction for video reasoning: language should organize and communicate predictions, while latent visual states preserve the dynamic semantics needed to imagine what happens next.

Limitations

FUTURE-L1 is developed for future-oriented video reasoning, where intermediate dynamic visual states are especially useful. Its SFT stage uses TwiFF-style trajectories with future reasoning frames; extending the same recipe to domains without such annotations may require automatic hint selection or teacher distillation. Our experiments focus on Qwen3-VL-8B and two VEP benchmarks, and future work can explore larger backbones, broader video tasks, and alternative sources of latent visual supervision. Furthermore, our evaluation data consist of instructional, predictive and English-language videos shot with static cameras, and generalization to multilingual, cinematic, or dynamically captured content remains open.

Societal Impact

FUTURE-L1 is designed for academic research on video event prediction. However, the ability to infer future behavior from video prefixes could be repurposed for surveillance systems, predictive policing, or behavior anticipation in public spaces. Erroneous predictions could lead to harmful outcomes, including false accusations or discriminatory profiling. Our current evaluation uses only publicly available research benchmarks with no real-world deployment. We encourage the community to consider responsible use guidelines when building upon future-oriented video reasoning methods.

Acknowledgments

We thank the Shanghai Artificial Intelligence Laboratory for their institutional support. This work was supported by the Shanghai Artificial Intelligence Laboratory under Project No. PJ-PRJ24DMX001.

References

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge,

Zimin Ran, and 11 others. 2026. [Llava-onevision-2: Towards next-generation perceptual intelligence](#). *Preprint*, arXiv:2605.25979.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Jen-Hao Cheng, Vivian Wang, Huayu Wang, Huapeng Zhou, Yi-Hao Peng, Hou-I Liu, Hsiang-Wei Huang, Kuang-Ming Chen, Cheng-Yen Yang, Wenhao Chai, et al. 2025a. Tempura: Temporal event masked prediction and understanding for reasoning in action. *arXiv preprint arXiv:2505.01583*.

Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. 2025b. Video-as-answer: Predict and generate next video event with joint-grpo. *arXiv preprint arXiv:2511.16669*.

Tao Cheng, Shi-Zhe Chen, Hao Zhang, Yixin Qin, Jinwen Luo, and Zheng Wei. 2026. Hylar: Hybrid latent reasoning with decoupled policy optimization. In *Computer Vision – ECCV 2026*, pages 613–630, Cham. Springer Nature Switzerland.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. 2025. [Emerging properties in unified multimodal pretraining](#). *Preprint*, arXiv:2505.14683.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. [Video-r1: Reinforcing video reasoning in mllms](#). *Preprint*, arXiv:2503.21776.

Chaoyou Fu, Yuhao Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. [Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis](#). *Preprint*, arXiv:2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su,

- Chengwu Long, Xiaoyao Xie, et al. 2026. Videomme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*.
- Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2019. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.
- Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, et al. 2026. Glm-5v-turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*.
- Tianxiang Jiang, Sheng Xia, Yicheng Xu, Linqun Wu, Xiangyu Zeng, Limin Wang, Yu Qiao, and Yi Wang. 2025. Vknowu: Evaluating visual knowledge understanding in multimodal llms. *arXiv preprint arXiv:2511.20272*.
- Hema S. Koppula and Ashutosh Saxena. 2016. [Anticipating human activities using object affordances for reactive robotic response](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29.
- Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. 2014. A hierarchical representation for future action prediction. In *Computer Vision – ECCV 2014*, pages 689–704, Cham. Springer International Publishing.
- Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8769–8784.
- Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Hao Chen, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2025a. Latent visual reasoning. *arXiv preprint arXiv:2509.24251*.
- Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. 2025b. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Songze Li, Zun Wang, Gengze Zhou, Jialu Li, Xiangyu Zeng, Limin Wang, Yu Qiao, Qi Wu, Mohit Bansal, and Yi Wang. 2025c. Learning goal-oriented language-guided navigation with self-improving demonstrations at scale. *arXiv preprint arXiv:2509.24910*.
- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025d. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.
- Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. 2025. [Videvent: A large dataset for understanding dynamic evolution of events in videos](#). *Preprint*, arXiv:2506.02448.
- Junhua Liu, Zhangcheng Wang, Zhike Han, Ningli Wang, Guotao Liang, and Kun Kuang. 2026a. Twiff (think with future frames): A large-scale dataset for dynamic visual reasoning. *arXiv preprint arXiv:2602.10675*.
- Naiming Liu, Zhiheng Wu, Shuning Wang, Tie Zhang, Bowen Liu, and Tong Wang. 2026b. [Ovibench: Benchmarking online video question answering under interruption](#). *Preprint*, arXiv:2608.22279.
- Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, et al. 2026c. Videoauto-r1: Video auto reasoning via thinking once, answering twice. *arXiv preprint arXiv:2601.05175*.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772.
- Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, et al. 2026a. Onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*.
- Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, Zhewen Hao, Xi Gao, Shaoheng Nie, Yixuan Wei, Zhenda Xie, Ting Chen, and Gang Zeng. 2026b. Thinking with visual primitives.
- OpenAI. 2024. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o>.
- Tan-Hanh Pham and Chris Ngo. 2025. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *arXiv preprint arXiv:2508.12587*.

- Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and Xudong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.
- MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. 2014. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*.
- Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 677–693.
- Yansong Shi, Qingsong Zhao, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. 2026. River: A real-time interaction benchmark for video llms. *arXiv preprint arXiv:2603.03985*.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. *Openai gpt-5 system card*. Preprint, arXiv:2601.03267.
- Qile Su, Jing Tang, Rui Chen, Lei Sun, and Xiangxiang Chu. 2026. Video-coe: Reinforcing video event prediction via chain of events. *arXiv preprint arXiv:2603.14935*.
- Qile Su, Shoutai Zhu, Shuai Zhang, Baoyu Liang, and Chao Tong. 2025. Eventformer: A node-graph hierarchical attention transformer for action-centric video event prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4698–4707.
- Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026a. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.
- V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026b. *Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning*. Preprint, arXiv:2507.01006.
- Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. 2025. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*.
- Jintao Tong, Shilin Yan, Hongwei Xue, Xiaojun Tang, Kunyu Shi, Guannan Zhang, Ruixuan Li, and Yixiong Zou. 2026. Swimbird: Eliciting switchable reasoning mode in hybrid autoregressive mllms. *arXiv preprint arXiv:2602.06040*.
- Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016a. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106.
- Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016b. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29.
- Chenting Wang, Kunchang Li, Tianxiang Jiang, Xiangyu Zeng, Yi Wang, and Limin Wang. 2025a. Make your training flexible: Towards deployment-efficient video models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23880–23891.
- Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. 2025b. Fostering video reasoning via next-event prediction. *arXiv preprint arXiv:2505.22457*.
- Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025c. *Videorf: Incentivizing video reasoning capability in mllms via reinforced fine-tuning*. Preprint, arXiv:2505.12434.
- Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. 2025d. Monet: Reasoning in latent visual space beyond images and language. *arXiv preprint arXiv:2511.21395*.
- Shuning Wang, Zhiheng Wu, YiNuo Lu, Naiming Liu, Chen Jia, Bowen Liu, Shuo Nie, Weijie Zhu, and Yumeng Zhang. 2026. See more, think deeper: Query-expanded visual evidence and answer-clue guided reflection for long video understanding. *arXiv preprint arXiv:2606.09064*.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025e. Internv13. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. 2024. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pages 396–416. Springer.
- Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. 2025. Sim-cot: Supervised implicit chain-of-thought. *arXiv preprint arXiv:2509.20317*.

- Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. 2026a. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*.
- Zhiheng Wu, Tong Wang, Shuning Wang, Naiming Liu, and Yumeng Zhang. 2026b. See further, think deeper: Advancing vlm’s reasoning ability with low-level visual cues and reflection. *arXiv preprint arXiv:2604.24339*.
- Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. 2026. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*.
- LLM-Core-Team Xiaomi. 2025. [Mimo-vl technical report](#). *Preprint*, arXiv:2506.03569.
- Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, et al. 2025. Expvid: A benchmark for experiment video understanding & reasoning. *arXiv preprint arXiv:2510.11606*.
- Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. 2026. Internvideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*.
- Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025a. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643.
- Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. 2025b. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). *Preprint*, arXiv:2503.14476.
- Xinlei Yu, Zhangquan Chen, Yongbo He, Tianyu Fu, Cheng Yang, Chengming Xu, Yue Ma, Xiaobin Hu, Zhe Cao, Jie Xu, et al. 2026. The latent space: Foundation, evolution, mechanism, ability, and outlook. *arXiv preprint arXiv:2604.02029*.
- Xiangyu Zeng, Zhiqiu Zhang, Yuhao Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. 2026. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. [LMMS-eval: Reality check on the evaluation of large multimodal models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025b. [Llava-video: Video instruction tuning with synthetic data](#). *Preprint*, arXiv:2410.02713.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
- Yaowei Zheng, Juntong Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. 2025a. Easyrl: An efficient, scalable, multimodality rl training framework. <https://github.com/hiyouga/EasyRL>.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025b. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv (Cornell University)*.
- Yubo Zhu, Dongrui Liu, Zecheng Lin, Wei Tong, Sheng Zhong, and Jing Shao. 2025b. The llm already knows: Estimating llm-perceived question difficulty via hidden representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1176.

A Baselines

General MLLMs. We compare against broadly trained open-source and proprietary multimodal models, including GLM-4.1V (Team et al., 2026b), LLaVA-NeXT-Video (Zhang et al., 2025b), MiMo-VL (Xiaomi, 2025), InternVL3 (Zhu et al., 2025a), Qwen2.5/3-VL (Bai et al., 2025b,a), GPT-4o (OpenAI, 2024), and GPT-5 (Singh et al., 2026). These models test whether generic video-language instruction following is sufficient for future-event prediction.

Video-Reasoning Models. We also include methods that explicitly train or optimize video reasoning behavior, including Video-RFT (Wang et al., 2025c), Video-R1 (Feng et al., 2025), VideoAuto-R1 (Liu et al., 2026c), Video-o3 (Zeng et al., 2026), NEP (Wang et al., 2025b), and Video-CoE (Su et al., 2026). Most of these baselines use SFT, RL, or both to strengthen textual reasoning over video; they are the closest text-centric competitors to our latent visual reasoning pipeline.

Latent Visual Reasoning Models. We also compare against LVR (Li et al., 2025a), Monet (Wang et al., 2025d), and SwimBird (Tong et al., 2026). These models introduce non-textual or latent visual reasoning mechanisms, but were primarily developed outside dense future-event prediction. Their transfer performance helps separate the benefit of latent reasoning in general from the specific data curation and latent-aware RL used by FUTURE-L1.

Unified Models. For TwiFF-Bench, we follow the benchmark protocol and compare against representative MLLMs (Qwen2.5-VL, InternVL3.5, and DeepEyes) as well as unified understanding-generation models. Janus-Pro (Chen et al., 2025) and Bagel (Deng et al., 2025) are unified multimodal models that support both visual understanding and generation, making them relevant baselines for future-frame reasoning beyond pure text QA. TwiFF-300K and TwiFF-2.7M (Liu et al., 2026a) are trained on large-scale interleaved future-frame reasoning data and therefore represent the strongest TwiFF-specific unified baselines. These comparisons evaluate both the quality of the generated reasoning trajectory and the correctness of the final open-ended answer.

Algorithm 1 LA-DAPO: one update step

Require: policy π_θ , prompt x , group size G , coefficients $\lambda_{a,f,c,d}$, temperature τ

- 1: Sample G rollouts $\{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | x)$; record latent trajectories $\{\mathbf{Z}_i\}$
- 2: **for** $i = 1, \dots, G$ **do**
- 3: $R_{\text{acc}}(i), R_{\text{fmt}}(i) \leftarrow \text{rule+judge / template match on } y_i$
- 4: $R_{\text{ctr}}(i) \leftarrow \text{hardest-positive InfoNCE on } \mathbf{Z}_i \text{ vs. group (Eq. 5)}$
- 5: $R_{\text{div}}(i) \leftarrow -\frac{1}{M-1} \sum_m \cos^2(\mathbf{b}_m, \mathbf{b}_{m+1})$ (Eq. 6)
- 6: $R_i \leftarrow \lambda_a R_{\text{acc}}(i) + \lambda_f R_{\text{fmt}}(i) + \lambda_c R_{\text{ctr}}(i) + \lambda_d R_{\text{div}}(i)$
- 7: **end for**
- 8: $\hat{A}_i \leftarrow (R_i - \text{mean}(R)) / \text{std}(R)$; broadcast to text/answer tokens
- 9: Update θ by maximizing $\mathcal{J}_{\text{DAPO}}$ with $\hat{A}_i$

B LA-DAPO Details

DAPO Objective. DAPO (Yu et al., 2025) optimizes a token-level surrogate with an asymmetric PPO clip. Let $r_t(\theta) = \pi_\theta(y_t | y_{<t}, x) / \pi_{\theta_{\text{old}}}(y_t | y_{<t}, x)$ be the importance ratio and $\hat{A}_t$ the group-relative advantage. The per-token loss is

$$\mathcal{J}_{\text{DAPO}} = \mathbb{E}[\min(r_t \hat{A}_t, \text{clip}(r_t, 1 - \epsilon_l, 1 + \epsilon_h) \hat{A}_t)],$$

with $\epsilon_h > \epsilon_l$ so positive-advantage tokens enjoy a wider exploration window than negative ones. We use $\epsilon_l=0.2$, $\epsilon_h=0.28$, $G=8$, and apply this loss only to textual and answer tokens.

Accuracy Reward $R_{\text{acc}} \in \{0, 1\}$. We extract the `<answer>...</answer>` span and judge it against the ground truth. Rule-based matching is tried first; the LLM judge (Qwen3.6-27B-Instruct) is invoked only on rule-failed samples, using the prompt in Figure 14. $R_{\text{acc}}=1$ if either step accepts the prediction, else 0.

Format Reward $R_{\text{fmt}} \in \{0, 1\}$. Following the template in Figure 3, a valid response opens with a `<reason>...</reason>` block, alternates one or more `<|latent_start|>...<|latent_end|>` latent spans each followed by a `<reason>...</reason>`, and terminates with a single `<answer>...</answer>`. Responses with zero latent spans (pure text CoT) or any malformed boundary receive $R_{\text{fmt}}=0$.

Edge Cases for R_{ctr} . Group accuracy is frequently skewed early in RL, so we define R_{ctr} explicitly when either partition is empty: when $\mathcal{P}_i=\emptyset$ we set $R_{\text{ctr}}(i) = 1/(1 + \sum_{j \in \mathcal{N}_i} \exp(s_{ij}/\tau))$, rewarding the absence of close negatives; when $\mathcal{N}_i=\emptyset$ we use s_i^+/τ ; when both are empty we set $R_{\text{ctr}}(i) = 0$.

Table 8: **SFT hyperparameters.** Settings used to train FUTURE-L1-SFT.

Item	Value
Initialization	Qwen3-VL-8B-Instruct (Bai et al., 2025a)
Training data	FUTURE-L1-50K
LLM Backbone	Full tuning
Vision tower / merger	Frozen
Precision	bf16
engine	DeepSpeed ZeRO-2
Optimizer	AdamW
β_1, β_2	0.9, 0.95
Weight decay	0.1
Gradient clip	1.0
Schedule / warm-up	Cosine / 0.1
Peak LR	1×10^{-5}
Global batch	128
Sequence length	16,384
Frames	16
MSE weight	$\lambda=0.1$
Latent budget	$L_{\max}=4$

Table 9: **RL / LA-DAPO hyperparameters.** Settings used to train FUTURE-L1-RL.

Item	Value
Initialization	FUTURE-L1-SFT checkpoint
Training data	FutureBench: 2K; TwiFF-Bench: 20K
RL framework	<i>Easy-RL</i> (Zheng et al., 2025a)
Rollout batch	64
Group size	$G=8$
Max prompt length	8,192
Max response length	2,048
Temperature / top- p	0.9/0.99
λ_a	0.9
λ_f	0.1
Clip	$\epsilon_l=0.2, \epsilon_h=0.28$
Dual clip	3.0
KL coeff.	10^{-2}
Group filter	mean acc. $\in [0.1, 0.9]$
Judge model	Qwen3.6-27B

Edge Cases for R_{div} . When a rollout produces $M \leq 1$ latent spans there are no adjacent pairs and we set $R_{\text{div}}=0$. Span representatives with zero L_2 norm are skipped to avoid division-by-zero in the cosine.

C Implementation Details

The training hyperparameters for the SFT and LA-DAPO stages are summarized in Tables 8 and 9, respectively. We implement the RL stage with the open-source *Easy-RL* framework (Zheng et al., 2025a).

D Additional Evaluation Details

D.1 Benchmark Details

FutureBench. FutureBench (Wang et al., 2025b) evaluates multiple-choice video event prediction from an observed video prefix. Each example provides a video, a question, four candidate future-event continuations, and a single correct option. The benchmark separates examples by temporal reasoning depth: 1-Hop asks for the next immediate future event, 2-Hop and 3-Hop require progressively longer event chains, and Interp. requires reasoning over non-consecutive future events under partial intermediate anchors. We report overall accuracy and the four split accuracies. For RL, we follow NEP (Wang et al., 2025b) and Video-CoE (Su et al., 2026) and train LA-DAPO for one

Twiff-Bench. Twiff-Bench (Liu et al., 2026a) evaluates open-ended future-frame reasoning. Each example contains input frames sampled from the observed prefix, a forecasting question, reference future reasoning with intermediate reasoning images, and a ground-truth answer. The task covers instructional, predictive, and camera-centric scenarios. Unlike FutureBench, Twiff-Bench is not a multiple-choice benchmark: it evaluates both the model’s reasoning trajectory and final answer on a 0–5 scale, and the reported score is the average of the two dimensions. For RL, we randomly sample 20K format-valid examples from the retained visual-gain pool and train for one epoch. All SFT and RL training sets are filtered to be disjoint from the reported benchmark evaluation sets, ensuring no overlap between training examples and measured test samples.

For FutureBench, we evaluate each sample with up to 32 input frames and allow at most 2,048 new tokens. For TwiFF-Bench, we allow at most 4,096 new tokens. Both benchmarks use deterministic decoding: temperature 0, top- p 1, beam size 1, and sampling disabled.

FUTURE-L1-50K is the 50K subset used to cold-start latent visual reasoning before LA-DAPO. It is selected from TwiFF-format interleaved trajectories by the visual-gain probe described in §3.2. Each example contains a video prefix frame, one or more future reasoning frames, and an interleaved textual reasoning trace. The retained examples emphasize cases where future visual hints substantially improve prediction reliability, so the dataset targets samples for which visual imagination is empirically useful rather than merely available.

Figure 1 consists of four bar charts showing the distribution of data across different categories and reasoning steps.

- Top-left chart:** Distribution by Category. The y-axis is '# examples' (0 to 20,000). The x-axis is 'Category' (Instructional, Predictive, Camera). The distribution is: Instructional (54.7%), Predictive (33.6%), Camera (11.7%).
- Top-right chart:** Distribution by Visual gain $p_r - p_l$. The y-axis is '# examples' (0 to 10,000). The x-axis is 'Visual gain $p_r - p_l$ ' (3 to 8). The distribution is: 3 (14.1K), 4 (11.3K), 5 (8.2K), 6 (6.3K), 7 (5.1K), 8 (5.1K).
- Bottom-left chart:** Distribution by # reasoning frames. The y-axis is '# examples' (0 to 30,000). The x-axis is '# reasoning frames' (1 to 7). The distribution is: 1 (31.3K), 2 (16.6K), 3 (1.7K).
- Bottom-right chart:** Distribution by Words per example. The y-axis is '# examples' (0 to 20,000). The x-axis is 'Words per example' (0 to 400). The distribution is: 100 (1.5K), 150 (20.5K), 200 (6.5K), 250 (1.5K). A dashed line indicates the mean = 128.8.

● Text
 ● Vision
 ● 1st Latent Span
 ● 2nd Latent Span
 ● 3rd Latent Span
 ● 4th Latent Span

Stage-wise Latent States. Figure 8 examines whether latent spans collapse to redundant states. We visualize token embeddings from FUTURE-L1-RL on FutureBench and group latent states by span order. Text and vision tokens occupy separate modality regions, while ordered latent spans form compact clusters that are also separated from

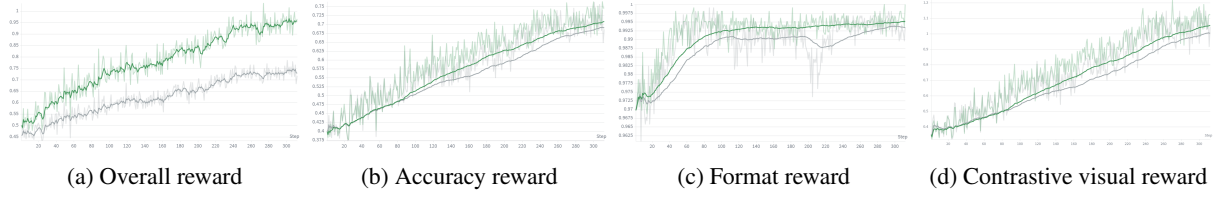


Figure 9: **Reward dynamics during RL.** FUTURE-L1 shows higher and more stable rewards than DAPO.

one another. This structure suggests that the model is not repeatedly emitting the same latent visual thought across time. Instead, the latent channel provides a stage-wise representation process in which successive spans update the model’s internal future hypothesis before the final prediction.

Reward Dynamics. Figure 9 compares the training rewards of standard DAPO and our latent-aware FUTURE-L1 policy. Across the overall reward, accuracy reward, format reward, and contrastive visual reward, FUTURE-L1 consistently yields higher and more stable trajectories than DAPO. The advantage is not limited to the final-answer signal: the contrastive visual reward also improves, indicating that LA-DAPO aligns latent visual states with successful prediction trajectories rather than merely optimizing textual answer format. These dynamics provide training-time evidence that the proposed latent-aware rewards make RL more effective for future-event reasoning.

G Prompts

Figure 10 shows the system prompt that enables interleaved textual and latent visual reasoning. For TwiFF-Bench evaluation, Figure 11 gives the user prompt template, while Figures 12 and 13 specify the judge prompt and payload used to score reasoning quality and answer accuracy. Figure 14 reports the binary answer-judge prompt used by the LA-DAPO accuracy reward.

H Case Study

Figures 15–17 provide successful qualitative examples on FutureBench. In these cases, FUTURE-L1 does not compress the whole forecast into a single textual chain. Instead, it alternates short verbal anchors with latent spans at points where the future state changes: entering a new room, manipulating an object, moving from a product setup to outdoor use, or transitioning across action stages. The textual tokens make the trajectory readable, while the latent spans mark intermediate visual hypotheses

that need to be carried forward before choosing the final option.

Figure 18 illustrates a representative failure. The model identifies the high-level baseball-dog context, but its latent trajectory drifts toward a plausible generic continuation and misses the specific ground-truth sequence involving the dog on the “BASEBALL” carpet, the open refrigerator, and the later dugout scene. This suggests that invoking latent spans is not sufficient by itself: the sampled latent trajectory must also preserve fine-grained event identity. This motivates the LA-DAPO stage, which optimizes latent trajectories with outcome-contrastive and temporal-diversity rewards.

I Checklist

I.1 Artifact Licenses and Terms

We use FutureBench, TwiFF-Bench, Qwen3-VL, *lmms-eval*, and *Easy-RI* under their respective licenses and terms. Any released code, checkpoints, or processed data will include explicit license information and will not redistribute third-party assets beyond permitted terms.

I.2 Artifact Intended Use

We use existing artifacts only for their intended research purposes: benchmark evaluation, model initialization, training, and academic comparison. Derived artifacts are intended for research use in video event prediction and multimodal reasoning, consistent with the original access conditions.

I.3 Artifact Documentation

The artifacts cover English-language video event prediction and multimodal reasoning. FutureBench provides multiple-choice forecasting splits, TwiFF-Bench provides open-ended future-frame reasoning examples, and Appendix E documents FUTURE-L1-50K statistics; we do not use or infer demographic attributes.

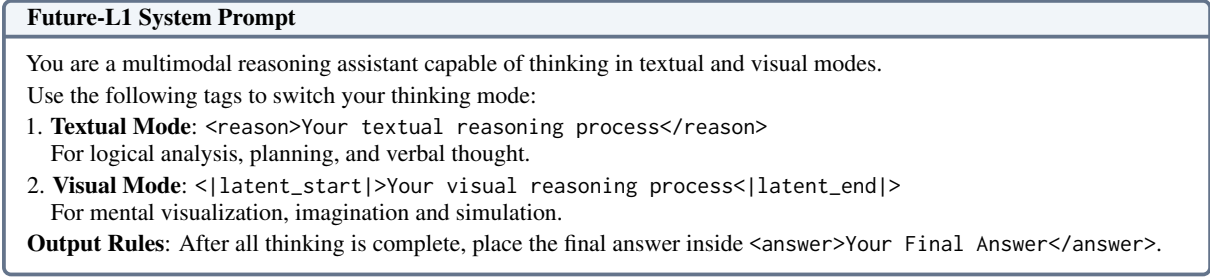


Figure 10: **Future-L1 system prompt.**

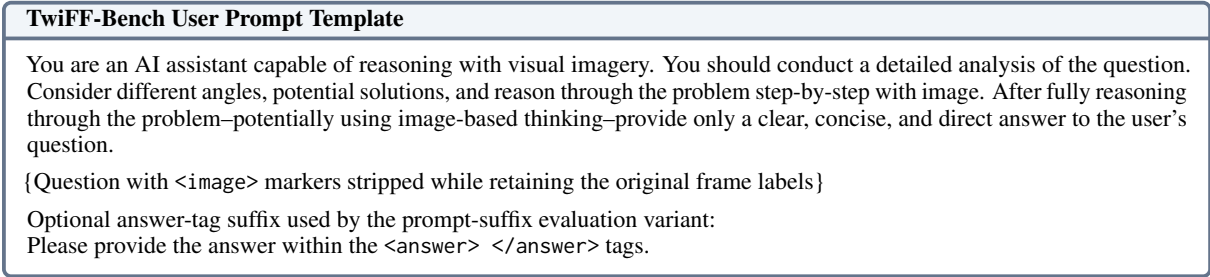


Figure 11: **Twiff-Bench user prompt template.**

I.4 Data Statistics

We report the main data statistics in the paper and appendix: FUTURE-L1-50K contains 50K examples, RL uses 2K FutureBench and 20K TwiFF-style training examples, TwiFF-Bench has 1,078 evaluation QA samples, and all training data are filtered to be disjoint from the reported evaluation sets.

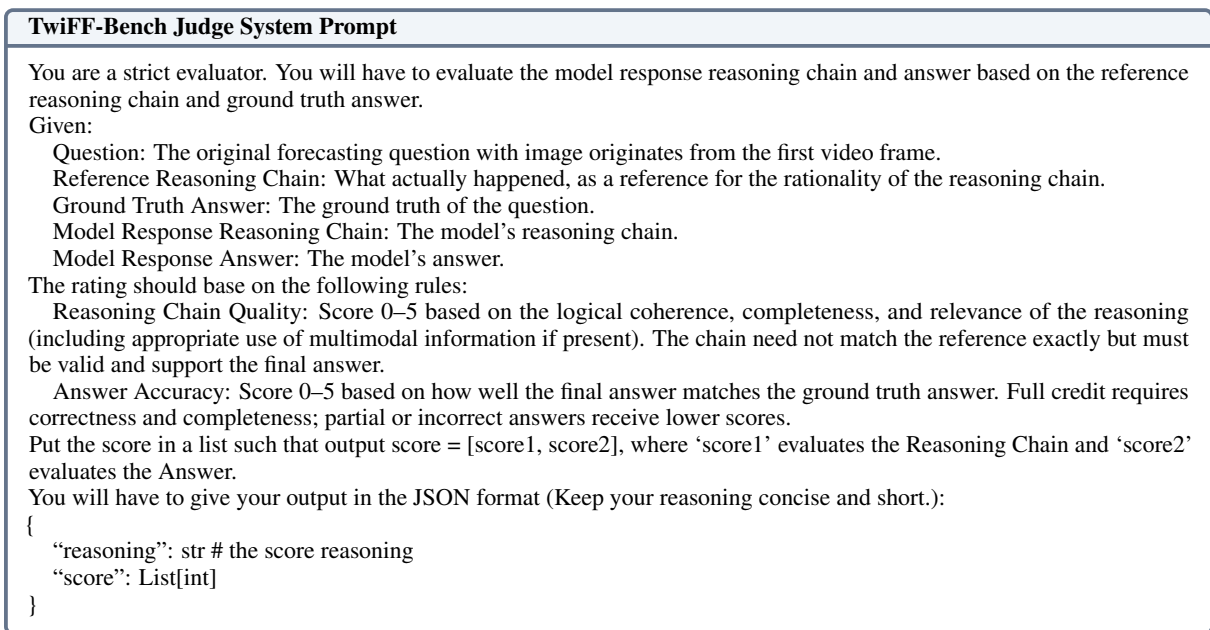


Figure 12: Twiff-Bench judge system prompt.

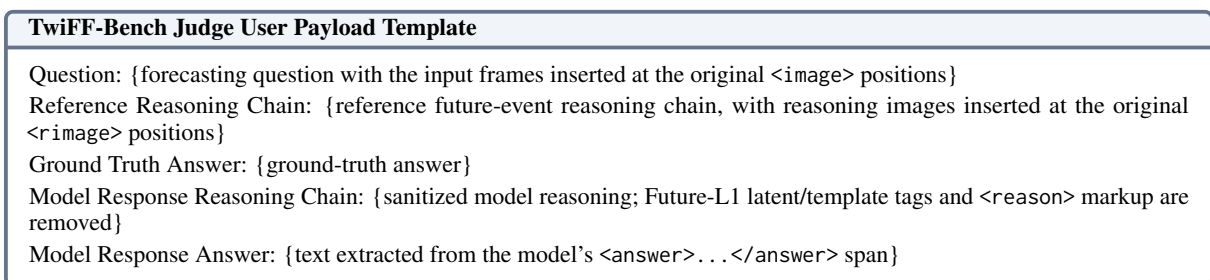


Figure 13: Twiff-Bench judge user payload template.

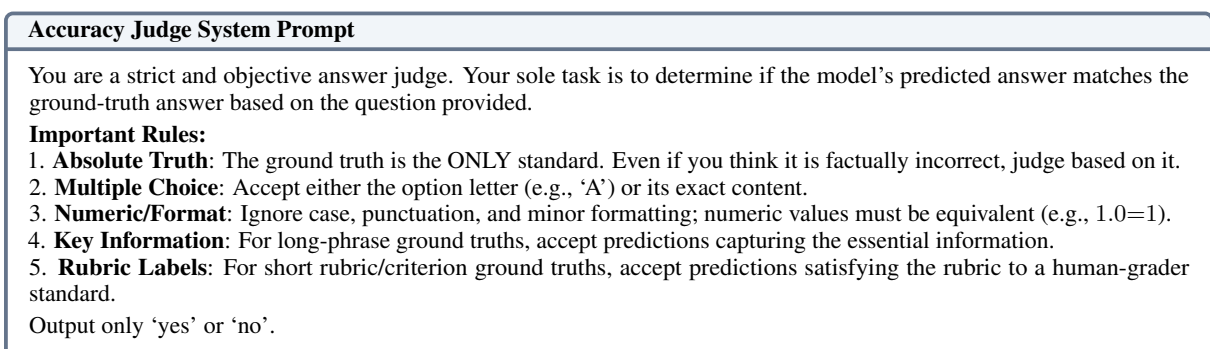


Figure 14: Accuracy judge system prompt.

Input Video



Question

Based on the given video, predict future events and fill in the potential events in the following sequence: 1. [?] 2. [?] 3. [?] 4. [A black screen displaying the text 'Your Face Music']. Which of the following options best fits the intermediate events between the observed scenes and the final scene?

Options

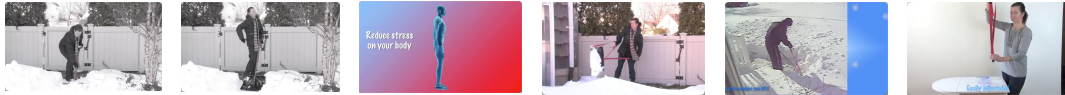
- A. 1. The person enters the bathroom and uses an electric trimmer to shave their beard. 2. They examine their partially shaved beard in the mirror and seal a clear plastic bag of trimmed hair. 3. They return to the dimly lit bedroom, lying on the bed in a contemplative mood while holding a small, dark-colored object.
- B. 1. The person returns to the bedroom and immediately lies in bed while holding the small object. 2. They then enter the kitchen to find a snack. 3. Afterwards, they use the trimmer in the bathroom to shave their beard.
- C. 1. The person goes into the bathroom to take a hot shower. 2. They then use the electric trimmer to shave their beard and rinse off in the sink. 3. Finally, they stand in the doorway of the bathroom, holding the dark object as if calling someone.
- D. 1. The person skips any grooming and instead checks a smartphone. 2. They leave the bedroom to make a phone call outdoors. 3. Later, they return and sit on the bed absentmindedly before the video ends with the black text.

Future-L1 interleaved Trajectory

<reason>The video begins with a person lying in bed, seemingly itchy, as indicated by the text 'FEELING ITCHY?'. This sets the context for a grooming routine.</reason>
<|latent_start|><|latent|>...<|latent|><|latent_end|>
<reason>The person then sits up and picks up a small, dark-colored object, which appears to be an electric trimmer. This suggests the intention to groom themselves.</reason>
<|latent_start|><|latent|>...<|latent|><|latent_end|>
<reason>The scene transitions to a bathroom where the person is seen using the trimmer to shave their beard, indicating the start of the grooming process.</reason>
<|latent_start|><|latent|>...<|latent|><|latent_end|>
<reason>After shaving, the person examines their partially shaved beard in the mirror and seals a clear plastic bag of trimmed hair, showing the completion of the grooming task.</reason>
<|latent_start|><|latent|>...<|latent|><|latent_end|>
<reason>The person then returns to the dimly lit bedroom, lying on the bed in a contemplative mood while holding the small, dark-colored object, which could be a phone or another personal item. This sequence of events aligns with the logical progression of grooming and the final scene of the video.</reason>
<answer>A</answer>

Figure 15: **Successful case: grooming routine.** From an observed bedroom scene, FUTURE-L1 predicts the missing sequence of beard trimming, mirror inspection, and returning to bed. The latent spans are inserted around scene and action transitions, while the text keeps the forecast interpretable.

Input Video



Question

Based on the given video, predict future events and fill in the potential events in the following sequence: 1. [?] 2. [?] 3. [?] 4. [A final outdoor demonstration with a professional endorsement]. Which option correctly orders the intermediate events that lead to the final scene described above?

Options

- A. 1. A person unpacks the SHOVEL HELPER and positions a black shovel on a white table, 2. The individual attaches metal clips and tightens the thumbwheel to secure the product, 3. The tool is shown in action outdoors as the person shovels snow near a phone booth while explaining its features, 4. A person in winter attire demonstrates the tool next to a white fence and evergreen trees and receives a chiropractor's endorsement before the scene fades to black.
- B. 1. The tool is demonstrated outdoors as the person shovels snow near a phone booth while explaining its features, 2. A person unpacks the SHOVEL HELPER and positions a black shovel on a white table, 3. The individual attaches metal clips and tightens the thumbwheel to secure the product, 4. A person in winter attire demonstrates the tool next to a white fence and evergreen trees and receives a chiropractor's endorsement before the scene fades to black.
- C. 1. A person in winter attire demonstrates the tool next to a white fence and evergreen trees and receives a chiropractor's endorsement before the scene fades to black, 2. A person unpacks the SHOVEL HELPER and positions a black shovel on a white table, 3. The individual attaches metal clips and tightens the thumbwheel to secure the product, 4. The tool is shown in action outdoors as the person shovels snow near a phone booth while explaining its features.
- D. 1. A person unpacks the SHOVEL HELPER and positions a black shovel on a white table, 2. The tool is shown in action outdoors as the person shovels snow near a phone booth while explaining its features, 3. The individual attaches metal clips and tightens the thumbwheel to secure the product, 4. A person in winter attire demonstrates the tool next to a white fence and evergreen trees and receives a chiropractor's endorsement before the scene fades to black.

Future-L1 interleaved Trajectory

<reason>The video begins with a person unpacking the SHOVEL HELPER and positioning a black shovel on a white table, as seen in the initial frames. This sets the stage for the instructional demonstration of the product.</reason>

<||latent_start|><||latent|>...<||latent|><||latent_end|>

<reason>Next, the individual attaches metal clips and tightens the thumbwheel to secure the product, as shown in the subsequent frames. This step is crucial for the proper setup of the tool.</reason>

<||latent_start|><||latent|>...<||latent|><||latent_end|>

<reason>Following this, the tool is demonstrated in action outdoors as the person shovels snow near a phone booth while explaining its features, highlighting its ease of use and effectiveness. This outdoor demonstration showcases the practical application of the SHOVEL HELPER.</reason>

<||latent_start|><||latent|>...<||latent|><||latent_end|>

<reason>Finally, a person in winter attire demonstrates the tool next to a white fence and evergreen trees and receives a chiropractor's endorsement before the scene fades to black, emphasizing the product's benefits and credibility.</reason>

<answer>A</answer>

Figure 16: **Successful case: product demonstration.** FUTURE-L1 tracks the SHOVEL HELPER demonstration from table setup to attachment, outdoor use, and endorsement. The interleaved trajectory separates physical manipulation from later usage scenes.

Input Video



Question

Based on the given video, predict future events and fill in the potential events in the given future events: 1. [?] 2. [?] 3. [?] 4. [Scene 19: a shirtless man meditating on a rocky cliff, representing calm introspection]. Which option best represents the correct order of events from the observed scenes to the final scene?

Options

- A. 1. A live-action sequence in front of a traditional Chinese building with a person in yellow martial arts attire performing dynamic staff movements; 2. Shirtless performers in yellow pants balancing bowls on their heads and hands during martial arts stances; 3. A close-up of a fist with a person in dark attire preparing for a challenge with the word 'DIFFICULTY' and a progress bar; 4. A shirtless man meditating on a rocky cliff.
- B. 1. A close-up of a fist with a person in dark attire preparing for a challenge with the word 'DIFFICULTY' and a progress bar; 2. A live-action sequence in front of a traditional Chinese building with a person in yellow martial arts attire performing dynamic staff movements; 3. Shirtless performers in yellow pants balancing bowls on their heads and hands during martial arts stances; 4. A shirtless man meditating on a rocky cliff.
- C. 1. Shirtless performers in yellow pants balancing bowls on their heads and hands during martial arts stances; 2. A live-action sequence in front of a traditional Chinese building with a person in yellow martial arts attire performing dynamic staff movements; 3. A close-up of a fist with a person in dark attire preparing for a challenge with the word 'DIFFICULTY' and a progress bar; 4. A shirtless man meditating on a rocky cliff.
- D. 1. A live-action sequence in front of a traditional Chinese building with a person in yellow martial arts attire performing dynamic staff movements; 2. A close-up of a fist with a person in dark attire preparing for a challenge with the word 'DIFFICULTY' and a progress bar; 3. Shirtless performers in yellow pants balancing bowls on their heads and hands during martial arts stances; 4. A shirtless man meditating on a rocky cliff.

Future-L1 interleaved Trajectory

<reason>The video begins with a live-action sequence in front of a traditional Chinese building, where a person in yellow martial arts attire performs dynamic staff movements, setting the stage for martial arts practice (Scene 1).</reason>
<||latent_start|><||latent|>...<||latent|><||latent_end|>
<reason>Next, shirtless performers in yellow pants are shown balancing bowls on their heads and hands during martial arts stances, emphasizing balance and control (Scene 2).</reason>
<||latent_start|><||latent|>...<||latent|><||latent_end|>
<reason>Following this, a close-up of a fist with a person in dark attire preparing for a challenge with the word 'DIFFICULTY' and a progress bar appears, indicating a focus on strength and preparation (Scene 3).</reason>
<||latent_start|><||latent|>...<||latent|><||latent_end|>
<reason>The video then transitions to a shirtless man meditating on a rocky cliff, representing calm introspection (Scene 19).</reason>
<answer>A</answer>

Figure 17: **Successful case: staged action sequence.** FUTURE-L1 follows a martial-arts montage through performance, balance practice, challenge preparation, and the final meditation scene. The latent spans help bridge visually distinct future stages before the final answer.

Input Video



Question

Based on the given video, predict future events and fill in the potential events in the upcoming segments:

1. A group of players in light blue jerseys gathers on a baseball field with the golden retriever walking alongside a man in khaki pants and a dark blue jacket holding a baseball, while empty bleachers and a 'CAROLINA' sign form the backdrop. 2. [?] 3. The golden retriever is shown sitting on the field with a blue cap in its mouth, positioned among players standing quietly during what appears to be the national anthem. 4. [?] 5. A man in a light blue shirt and shorts is seen seated on the floor of an indoor room, happily playing with a colorful ball together with the golden retriever, while individuals in khaki pants and dark shirts can be seen in the background with tables and equipment. Which pair of missing scenes best completes the future events order?

Options

- A. Golden retriever lying on 'BASEBALL' carpet with man in black shirt; then, retriever exploring an open refrigerator and later in the dugout with a person in black shirt.
- B. Retriever exploring an open refrigerator and later in the dugout with a person in black shirt; then, golden retriever lying on 'BASEBALL' carpet with man in black shirt.
- C. Players giving fist bumps to the golden retriever on the field; then, man in a light blue shirt interacting with the dog in an indoor room with a 'BASEBALL' carpet.
- D. Man in a light blue shirt playing fetch on the baseball field; then, golden retriever lying on its back on a 'BASEBALL' carpet with a man in a black shirt.

Future-L1 interleaved Trajectory

<reason>The video begins with a golden retriever wearing a blue vest with 'UNC' and 'BASEBALL' written on it, walking through a hallway with people in blue shirts. This sets the stage for the dog's role in a baseball setting.</reason>

<|latent_start|><|latent|>...<|latent|><|latent_end|>

<reason>As the dog moves into a baseball dugout, it is seen interacting with players and a man holding a bucket, indicating its presence during team activities. The dog is then shown on the field, running and playing with a ball, highlighting its playful nature. The dog is later seen sitting on the field with a blue cap in its mouth, positioned among players standing quietly, suggesting a moment of focus or anticipation. The final scene shows a man in a light blue shirt and shorts playing with the dog on a 'BASEBALL' carpet, emphasizing the dog's role in team bonding and relaxation.</reason>

<answer>C</answer>

Figure 18: **Failure case: event-specific detail loss.** FUTURE-L1 recognizes the baseball-dog setting but predicts a generic continuation rather than the ground-truth sequence with the carpet, refrigerator, and dugout events. The example shows that latent invocation must still preserve fine-grained visual event identity.