WorldAgent: World Understanding via Physical Events

Anonymous ACL submission

Abstract

Physical world understanding refers to the ability to infer physical properties and predict object and environment dynamics from visual inputs, typically videos, and is essential for autonomous systems, robotics, and embodied agents. Although Vision–Language Models (VLMs) have made substantial progress in general video understanding, their ability to reason about physical dynamics remains limited. In this work, we introduce **WorldAgent**, a training-free agentic framework that enhances the physical reasoning capabilities of existing VLMs. Rather than processing videos as uniform sequences of frames, WorldAgent decomposes them into salient physical events. It is equipped with a physical event memory tool and a dynamics simulation tool to support perceptual, counterfactual, and predictive reasoning. We further propose an event-centric sampling strategy based on surprise scores from joint-embedding predictive architectures, which selects frames that capture informative and surprising physical moments, thereby significantly enhancing the effectiveness of both tools. Extensive experiments show that WorldAgent outperforms state-of-the-art VLMs (e.g., GPT-5.2) by up to 10% across three physical reasoning benchmarks. Moreover, event-centric sampling also improves long-video understanding, highlighting its potential as a general-purpose video sampling strategy. The code will be publicly available.

1 Introduction

Physical world understanding is the ability to look at a scene and infer its underlying physics, predicting how objects move or interact without performing explicit calculations (Battaglia et al., 2013; McCloskey, 1983). For embodied agents, this intuitive physics allows them to plan actions and navigate safely in real-world environments (Srivastava et al., 2022; Das et al., 2018). However, although modern vision-language models (Bai et al.; OpenAI, 2025; Bai et al.; Team et al., 2023) describe videos well and answer surface-level questions, they still struggle to match human intuition when understanding physical interactions in a scene (Chow et al., 2025; Bordes et al., 2025). This gap indicates that recognizing surface-level appearances fundamentally differs from inferring hidden (latent) dynamics. Prior studies (Chow et al., 2025; Balazadeh et al., 2025) attribute this limitation to insufficient physics-specific training in existing VLMs, which results in limited physical knowledge for robust physical understanding. In contrast, we argue that current VLMs already encode latent physical priors without extra training, but standard inference pipelines fail to access and exploit them effectively.

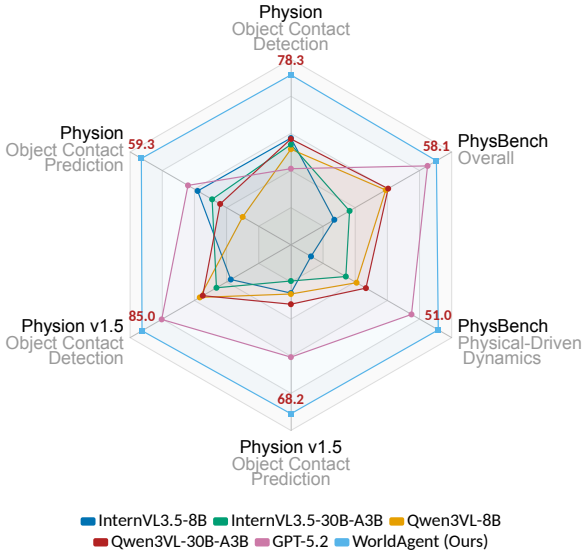

Figure 1: **WorldAgent Peformance.** It surpasses state-of-the-art closed-source (e.g., GPT-5.2) and open-source models (Qwen3VL, InternVL3.5) across physical world understanding benchmarks.

A key reason is that most VLMs process raw frame sequences end-to-end and lack an explicit intermediate stage that transforms observations into structured, event-centric physical variables (e.g., contacts, collisions, motion changes). In contrast, human physical reasoning operates in an inherently

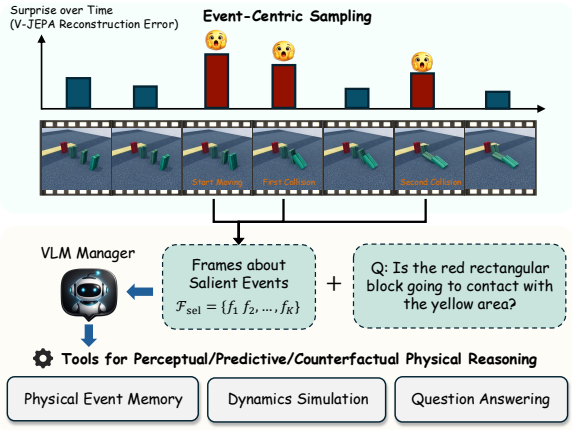

Figure 2: **Framework overview of WorldAgent.** WorldAgent first uses V-JEPA surprise scores to sample frames capturing salient physical events, such as motion onset and collisions. A VLM manager then selects and invokes either the physical event memory or the dynamics simulation tool according to the reasoning requirements of the physical question.

event-centric manner: humans extract a small number of salient physical events from a video and ground intuitive physical reasoning on these cues. For example, when a rolling ball collides with an obstacle, humans infer when and how the ball will bounce by focusing on the instant of contact rather than on intervals of steady motion. Inspired by this, we investigate whether adopting an event-centric reasoning paradigm can unlock the latent physical understanding already present in existing VLMs.

To achieve this, we introduce WorldAgent, an agentic framework that enables event-centric physical reasoning through two complementary agentic tools. The first tool, Physical Event Memory, extracts salient physical events and explains them using intuitive physical knowledge. It captures first-order dynamics through pairwise frame comparisons and then re-examines the video to summarize these interactions and model higher-order physical dynamics. This process enables WorldAgent to resolve complex perceptual reasoning tasks involving physical actions and phenomena. Beyond perceptual reasoning, physical understanding also requires predictive and counterfactual capabilities. To address this need, we introduce the Dynamics Simulation Tool, which leverages the extracted physical priors to simulate plausible outcomes under different contexts. By applying intuitive physical laws to the observed dynamics, the agent can mentally "rehearse" future or counterfactual scenarios and answer complex, high-level physical reasoning questions.

The efficacy of these agentic tools depends on visual inputs that accurately capture critical physical events. To address this, we propose an event-centric sampling strategy guided by the "surprise score" from the Video Joint Embedding Predictive Architecture (V-JEPA). Although prior work has used this signal to evaluate physical plausibility, we show that this score also serves as a robust proxy for event saliency. As shown in Fig. 2, frames with high surprise scores correlate with significant physical transitions, which ensures that the model focuses on the most informative dynamics.

As illustrated in Fig. 1, WorldAgent achieves SOTA results on Physion, Physion v1.5, and PhysBench, surpassing existing models by a wide margin. Our backbone-agnostic framework shows that event-level reasoning effectively leverages the latent physical priors encoded in modern VLMs.

In summary, our contributions are as follows:

- We **propose** WorldAgent, a novel agentic framework that shifts from uniform frame processing to event-centric physical reasoning, more closely mimicking human cognitive strategies.
- We **introduce** a dual-tool architecture comprising Physical Event Memory and Dynamics Simulation, enabling VLMs to perform complex perceptual, predictive, and counterfactual reasoning.
- We **develop** an event-centric sampling strategy utilizing V-JEPA surprise scores, demonstrating that these signals serve as a robust proxy for identifying salient physical transitions in video.
- We **achieve** SOTA results across three benchmarks, outperforming GPT-5.2 by up to 10% while running 2–3× faster than existing agents.

2 Related Work

2.1 Physical Understanding Models

Traditional physical world understanding models can be broadly categorized into three groups. The first group consists of physics-specialized models (Guen and Thome, 2020; Duan et al., 2022), which are explicitly designed to model physical dynamics, such as predicting future object states or trajectories. These models typically rely on task-specific supervision and are optimized for narrow objectives (e.g., next-state prediction or stability estimation). The second group comprises physical oracle models (Zheng et al., 2024; Tung et al., 2023), in which predefined physical rules or structured representations are incorporated to reason

about object interactions. While effective in constrained settings, these approaches depend heavily on handcrafted assumptions or are constrained by the quality of external object detectors. The third group includes joint-embedding predictive architectures (JEPAs) (Bardes et al., 2023; Assran et al., 2025), where future latent representations are learned. These methods have demonstrated strong performance in distinguishing physically plausible from implausible scenarios, but they do not directly support language-based interaction.

As a result, all of these methods are ill-suited for open-ended physical reasoning, particularly in scenarios that require flexible, language-based responses or generalization beyond predefined rule sets. Our WorldAgent is a training-free framework that preserves the language reasoning capabilities of existing vision–language models while enabling effective physical reasoning in diverse simulated and real-world environments.

2.2 Vision-Language Models

Vision–language models (VLMs) integrate visual perception with language understanding to enable unified multimodal reasoning and have become a central paradigm in modern AI. Recent large-scale VLMs, such as Qwen3-VL (Bai et al.), GPT-5 (OpenAI, 2025), Claude-4.5 (Anthropic, 2025), Gemini-3 Pro (Google, 2025), and InternVL-3.5 (Wang et al., 2025a), demonstrate strong performance across a wide range of vision–language applications, including image and video captioning, visual question answering, visual grounding, and spatial understanding. Beyond standalone VLMs, several works (Fan et al., 2024; Wang et al., 2024, 2025b; Chen et al., 2025; Yang et al., 2025; Liu et al., 2025a; Jang et al., 2024) have proposed agentic frameworks in which external tools and iterative VLM invocations are leveraged to tackle more complex tasks, such as long video understanding. However, empirical studies show that even state-of-the-art VLMs and agent-based methods perform substantially below human level on physical world understanding tasks (Chow et al., 2025; Ma et al., 2025). WorldAgent significantly enhances the physical world understanding capabilities of VLMs and achieves substantially higher efficiency than existing agent methods.

2.3 Benchmarks and Datasets

Numerous benchmarks (Ma et al., 2025; Jassim et al., 2023) have been proposed to evaluate physical world understanding. Early datasets such as IntPhys (Riochet et al., 2020) and IntPhys2 (Bordes et al., 2025) assess intuitive physics by distinguishing physically plausible from implausible videos via real–fake comparisons, but this paired formulation does not align with single-video reasoning and is ill-suited for evaluating vision–language models. Simulation-based datasets such as CLEVR (Johnson et al., 2017), CLEVRER (Yi et al., 2020), and CRAFT (Ates et al., 2022) cover limited physical scenarios and have largely been solved by modern general-purpose VLMs. More recent benchmarks, including Physion and Physion v1.5 (Bear et al., 2021), evaluate models on recognizing or predicting object collisions using physical properties and dynamics in diverse scenarios. However, these datasets predate the LLM era and were not originally designed as question-answering (QA) tasks. We therefore reformulate them into a multiple-choice QA format to enable VLM-friendly evaluation. Additionally, we evaluate on PhysBench (Chow et al., 2025), the largest and most recent physical reasoning benchmark, to provide a comprehensive assessment.

3 WorldAgent

In Fig. 3, we illustrate the overall workflow of WorldAgent. Given a query, a VLM manager first selects an appropriate reasoning pathway: for questions answerable via perceptual physical cues, it constructs and queries a Physical Event Memory, whereas for questions requiring predictive or counterfactual reasoning, it invokes the Dynamics Simulation module to infer plausible state transitions under the queried context. Both tools operate on frames selected through event-centric sampling, and the system produces the final answer by reasoning over either the constructed memory or the simulated outcomes. We detail this dual-tool architecture in Section 3.1 and introduce the event-centric sampling strategy in Section 3.2.

3.1 Dual-Tool Architecture

Physical Event Memory. Given a VLM ϕ, we first classify the question q into reasoning types. If q receives the perceptual label, ϕ invokes an event-extraction tool to build a *physical event memory* from sampled frames. Concretely, given selected frames $\mathcal{F}_{\text{sel}} = \{f_i\}_{i=1}^{K}$, ϕ analyzes consecutive frame pairs to extract local state changes. For each

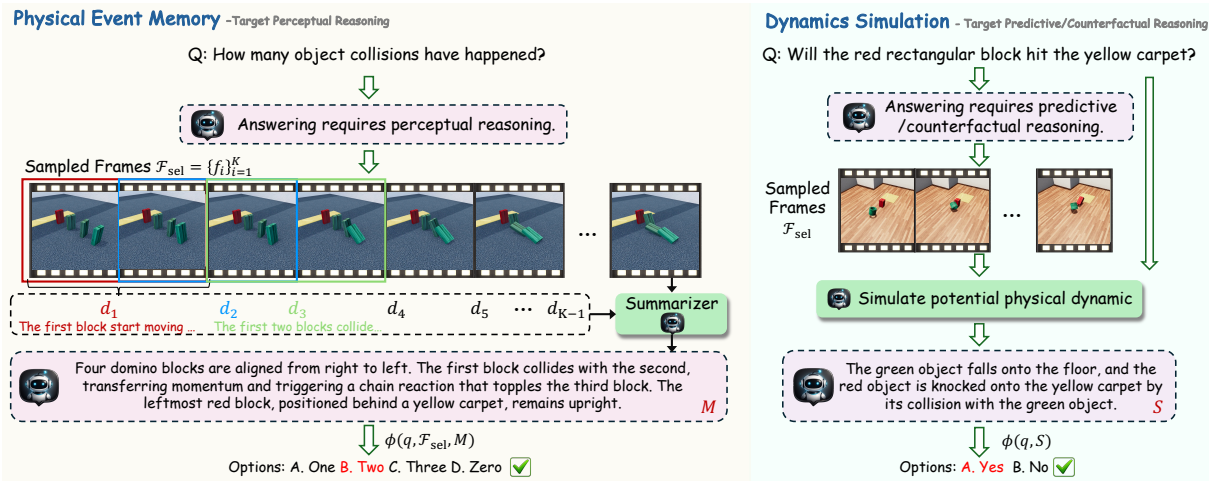

Figure 3: **Workflow overview of WorldAgent.** WorldAgent first samples K event-centric frames from a physical video. In parallel, the agent identifies the reasoning skill required to answer the physical question. If the question only requires perceptual reasoning (e.g., the number of collisions in the video), the agent constructs a physical event memory in two steps: (1) comparing physical state changes between frames and inferring their underlying causes, and (2) aggregating these comparisons into a coherent event summary. If the question requires counterfactual or predictive reasoning (e.g., the object state after the video ends), the agent simulates plausible outcomes conditioned on the frames. Finally, the agent uses the constructed memory or the simulation outcome to generate the answer.

pair (f_i, f_{i+1}), ϕ generates a transition description.

$$d_i = \phi(f_i, f_{i+1}), \quad i = 1, \ldots, K-1, \quad (1)$$

which captures the observed state change and its underlying physical cause. We explicitly prompt the model to infer causes using intuitive physical principles, such as momentum transfer during collisions and frictional resistance, rather than relying on purely commonsense descriptions. Each d_i follows the template: *"Frame i to i+1: state change description; this is because physical cause."*

The model ϕ then summarizes the resulting set of transition descriptions $\mathcal{D} = \{d_i\}_{i=1}^{K-1}$ into a compact physical event memory $M = \phi(\mathcal{D})$, which encodes the global temporal dynamics of $\mathcal{F}_{\text{sel}}$. This summarization filters out minor or redundant state changes and allows M to capture higher-order dynamics (e.g., acceleration) beyond first-order pairwise transitions. During this process, the model receives both the full video context and the transition descriptions $\mathcal{D}$ to guide the summarization. To prevent over-specification, we constrain the final memory length to fewer than 100 words, since longer memories do not consistently improve physical reasoning performance. See Appendix. B.2 for all prompts used for memory construction.

Finally, the model infers the answer by jointly conditioning ϕ on the question q, the sampled frames $\mathcal{F}_{\text{sel}}$, and the physical event memory M:

$$a_{mem} = \phi(q, \mathcal{F}_{\text{sel}}, M). \quad (2)$$

Dynamics Simulation. While the physical event memory suffices for perceptual physical reasoning, it becomes inadequate when a question requires extrapolating unobserved future states or reasoning about counterfactual physical outcomes. In such cases, WorldAgent invokes a *dynamics simulation* tool, which enables explicit reasoning over hypothetical state evolutions conditioned on the query and the observed video evidence.

Concretely, the VLM manager ϕ first assigns a reasoning label to the input question q. If it classifies the question as *predictive reasoning*, ϕ generates a simulation outcome that describes future state transitions beyond the temporal extent of the observed frames. If it classifies the question as *counterfactual reasoning*, ϕ instead simulates alternative outcomes under hypothetical interventions specified in the question. Both simulation modes require ϕ to first infer object-level physical properties from the sampled frames $\mathcal{F}_{\text{sel}}$, including attributes such as brittleness, stiffness, elasticity, softness, and friction. Conditioned on the inferred properties and the frames $\mathcal{F}_{\text{sel}}$, ϕ then produces a simulated description S, structured as a coherent sequence of plausible physical events.

Empirically, we find that reintroducing the original video frames during the final question-answering stage increases computational overhead and encourages ϕ to over-rely on perceptual cues, which weakens the effect of the simulated dynamics and degrades performance. We therefore condi-

Algorithm 1 Event-Centric Sampling

Require: Video frames $\{x_t\}_{t=1}^{T}$, sample size M, budget K, context encoder f_θ, target encoder f_ξ, predictor g_ϕ, mask tokens $U_{\mathcal{M}}$

Ensure: Sampled frames $\mathcal{F}_{\text{sel}}$

Uniform sampling M frames first.

$\{x_1, \ldots, x_M\} \leftarrow \textsc{UniformSamp}(\{x_t\}, M)$

Convert frames into context and target video clips.

$\mathcal{V} \leftarrow \{\underbrace{(x_1, x_2)}_{c_1^{ctx}}, \underbrace{(x_3, x_4)}_{c_1^{tgt}}, \ldots, \underbrace{(x_{M-1}, x_M)}_{c_{M/4}^{tgt}}\}$

Extract context and target embeddings.

$\{z_i^{\text{ctx}}\}, \{z_i^{\text{tgt}}\} \leftarrow f_\theta(\{c_i^{ctx}\}), f_\xi(\{c_i^{tgt}\})$

Calculate predicted target embeddings.

$\{\hat{z}_i^{\text{tgt}}\} = g_\phi\big([\,\{z_t^{\text{ctx}}\}; U_{\mathcal{M}}\,]\big)$

Compute surprise score.

$s_i \leftarrow \|\hat{z}_i^{\text{tgt}} - z_i^{\text{tgt}}\|_1 \;\; \forall i$

$\mathcal{V}_{\text{sel}} = f\{\, t_i \mid i \in \text{TopK}(\{s_i\})\,\}.$

Select event-centric frames.

$\mathcal{F}_{\text{sel}} \leftarrow \mathcal{V}_{\text{sel}}$

return $\mathcal{F}_{\text{sel}}$

tion the final answer prediction solely on the simulation outcome S, as follows:

$$a_{sim} = \phi(q, S). \tag{3}$$

3.2 Event-Centric Sampling

The effectiveness of the two agentic tools described above critically depends on the quality of the sampled frames. WorldAgent adopts an event-centric sampling strategy based on the *surprise score* introduced in V-JEPA2 (Assran et al., 2025). Accordingly, we first formally define this score.

V-JEPA Surprise Score Definition. Let a video be represented as a sequence of frames $\{x_t\}_{t=1}^{T}$. A shared visual backbone maps each frame x_t into a latent representation. The model samples a subset of frame indices $\mathcal{M} \subseteq \{1, \ldots, T\}$ as *masked targets* and treats the remaining indices $\mathcal{V} = \{1, \ldots, T\} \setminus \mathcal{M}$ as the *visible context*. A context encoder f_θ encodes the visible frames $\{x_t\}_{t\in\mathcal{V}}$ to produce context embeddings $\{z_t^{\text{ctx}}\}_{t\in\mathcal{V}}$, while a *target encoder* f_ξ encodes the masked frames $\{x_t\}_{t\in\mathcal{M}}$ to obtain the corresponding ground-truth target embeddings $\{z_t^{\text{tgt}}\}_{t\in\mathcal{M}}$.

The predictor network g_ϕ operates on the concatenation of the context embeddings $\{z_t^{\text{ctx}}\}_{t\in\mathcal{V}}$ and learnable mask tokens $U_{\mathcal{M}}$. These tokens specify the patch positions of the masked frames and enable the predictor to produce masked targets.

$$\{\hat{z}_i^{\text{tgt}}\} = g_\phi\big([\,\{z_t^{\text{ctx}}\}_{t\in\mathcal{V}}; U_{\mathcal{M}}\,]\big)\,. \tag{4}$$

The surprise score for each masked frame is defined as the discrepancy between the predicted and ground-truth target embeddings:

$$s_t = d\Big(\hat{z}_t^{\text{tgt}},\, z_t^{\text{tgt}}\Big), \qquad t \in \mathcal{M}, \tag{5}$$

where $d(\cdot,\cdot)$ denotes a distance metric.

Under the violation-of-expectation framework (Garrido et al., 2025), prior work shows that aggregating the surprise score over a video correlates with physical plausibility, where higher values indicate a greater likelihood of physical law violations. We further show that surprise scores at individual timesteps are also informative: frames with high scores capture salient events, whereas frames with low scores reflect predictable dynamics. Based on this, we design an event-centric sampling strategy driven directly by the surprise score. Unlike methods that rely on CLIP (Radford et al., 2021) similarity (Yao et al., 2025), optical flow (Liu et al., 2025b), or captions (Fan et al., 2024), our method is the first to use JEPA-based score as the primary sampling criterion.

Sampling Strategy Design. Technically, Algorithm 1 summarizes the implementation of our event-centric sampling procedure. Given a video $\{x_t\}_{t=1}^{T}$, we first uniformly sample M frames in temporal order and group them into $M/2$ non-overlapping two-frame clips $v_i = (x_{2i-1}, x_{2i})$. We then split these clips via interleaving assignment into a set of context clips $\{c_i^{\text{ctx}}\}_{i=1}^{M/4}$ and a set of target clips $\{c_i^{\text{tgt}}\}_{i=1}^{M/4}$. The context encoder f_θ jointly encodes all context clips to produce context embeddings $\{z_i^{\text{ctx}}\}$, which the predictor g_ϕ uses to estimate the latent representations of all target clips $\{\hat{z}_i^{\text{tgt}}\}$. The target encoder f_ξ encodes the target clips to obtain the corresponding target embeddings $\{z_i^{\text{tgt}}\}$. For each target clip, we compute a surprise score s_i as the ℓ_1 distance between $\hat{z}_i^{\text{tgt}}$ and z_i^{tgt}, then select the clips with the highest scores:

$$\mathcal{V}_{\text{sel}} = \{\, t_i \mid i \in \text{TopK}(\{s_i\})\,\}.$$

Finally, we output the first frame of each selected target clip as the event-centric sampled frames $\mathcal{F}_{\text{sel}}$. Unlike prior video agents that rely on iterative refinement, event-centric sampling operates in a single pass, making it more computationally efficient.

Table 1: Comparison of WorldAgent with general-purpose VLMs, Video VLMs, and agent-based methods on the Physion and Physion v1.5 benchmarks. Each benchmark includes two tasks: Object Contact Detection (OCD) and Object Contact Prediction (OCP). Our models are highlighted in blue.

Method	Physion		Physion v1.5	
	OCD	OCP	OCD	OCP
General-Purpose VLMs				
InternVL3.5-8B	71.5	55.8	70.1	55.2
Qwen3VL-8B	70.3	53.0	75.3	55.3
InternVL3.5-30B-A3B	70.8	54.9	72.5	53.9
Qwen3VL-30B-A3B	71.4	54.4	74.8	56.4
GPT-5.2	68.2	56.4	81.7	62.1
Video VLMs				
VideoLLaMA3	58.4	51.0	66.1	53.5
Video-LLaVA	51.7	50.3	22.7	27.6
VLM-Based Agents				
VideoTree	61.8	50.6	62.2	50.8
WorldAgent (Qwen3VL-8B)	74.1	58.8	79.6	62.4
WorldAgent (GPT-5.2)	78.3	59.3	85.0	68.2

4 Results

4.1 Evaluation Protocol

Metrics. We use Exact Match (EM) as the evaluation metric across all experiments. EM computes the proportion of predictions that exactly match the ground-truth answers. This metric is appropriate because all evaluation benchmarks comprise only closed multiple-choice questions.

Datasets. We evaluate on three established benchmarks for physical reasoning: Physion, Physion v1.5 (Bear et al., 2021), and PhysBench (Chow et al., 2025). Physion is a simulated video dataset comprising 1.2K video–question–answer triplets and is designed to assess intuitive physics understanding across eight simulated physical scenarios (e.g., rolling, sliding, and collisions). Each example asks: `Will the red object make contact with the yellow area?`, with a binary (yes/no) answer. Physion v1.5 follows the same task formulation as Physion while improving rendering quality and lighting diversity for more robust evaluation. Both Physion and Physion v1.5 include two tasks: Object Contact Detection (OCD), where the full video is provided, and Object Contact Prediction (OCP), where only the first 1.5 seconds are observed, requiring models to reason about future dynamics. Therefore, We utilize physical event memory for OCD, while OCP tasks are handled by a dynamics simulation tool. PhysBench consists of 10,002 physical reasoning questions spanning four domains: physical object properties, object relationships, scene understanding, and physics-driven dynamics. All questions are multiple choice with four options, and the options are either text or images. Answering these questions requires both strong physical knowledge and diverse reasoning capabilities, including perceptual, counterfactual, and predictive reasoning.

Baselines. We compare WorldAgent against three categories of models. First, general-purpose VLMs (Qwen3VL-8B, Qwen3VL-30B-A3B (Bai et al.), InternVL3.5-8B, InternVL3.5-30B-A3B (Wang et al., 2025a), GPT-5.2 (OpenAI, 2025)) are evaluated, which process videos as sequences of frames. Second, video VLMs (PLLaVA-7B (Xu et al., 2024), Video-LLaVA (Lin et al., 2024), Video-LLaMA3 (Zhang et al., 2025)) are evaluated, which are specifically optimized for video understanding. Third, the VLM-based agent VideoTree (Wang et al., 2025b) is evaluated, which is a training-free framework for long video understanding. Video-LLaMA3 is evaluated exclusively on the Physion benchmarks, since PhysBench requires interleaving reference frames during inference, which is not supported by Video-LLaMA3. We evaluate WorldAgent with two VLM backbones, Qwen3VL-8B and GPT-5.2. Since PhysBench contains 3,985 out of 10,002 image-only questions, WorldAgent directly uses predictions from its VLM backbone for these cases and applies the agentic framework only to video-based questions. As a result, performance gains on PhysBench may appear less pronounced.

4.2 Comparison with State-of-the-art

Table 1 reports EM performance on Physion and Physion v1.5 under both the OCD and OCP settings. With the exception of GPT-5.2, all comparison baselines struggle on the OCP task and achieve performance close to random chance. This result indicates that predicting future physical outcomes is substantially more challenging than recognizing previously observed events for current VLMs. Among the baselines, general-purpose VLMs consistently outperform video VLMs and other agent-based methods. We hypothesize that this behavior arises because most video VLMs are optimized for long-form video understanding, whereas physical reasoning videos in Physion are short (5–15 seconds) and require fine-grained temporal and event-level reasoning. We also observe weaker instruction-following in video VLMs, especially Video-LLaVA, which performs far below random chance on Physion v1.5 OCP.

Against this challenging backdrop, WorldAgent

Table 2: Comparison of WorldAgent with general-purpose VLMs, video VLMs, and agent-based methods on the PhysBench benchmark. The evaluation covers four categories: Object Properties (Property), Object Relationships (Relationships), Scene Understanding (Scene), and Physics-Based Dynamics (Dynamics). Our models are highlighted in blue. The best performance is marked in **bold**. Note: WorldAgent uses its backbone VLM for image-only questions and applies agentic tools only to video questions, which account for 6,017/10,002 samples.

Method	Property	Relationships	Scene	Dynamics	Average
Human	97.10	95.67	94.91	95.68	95.87
InternVL3.5-8B	53.8	54.7	32.4	41.5	45.4
Qwen3VL-8B	61.5	62.1	40.5	44.9	51.8
InternVL3.5-30B-A3B	56.5	58.7	30.5	44.1	47.3
Qwen3VL-30B-A3B	62.8	64.1	37.5	45.6	52.1
GPT-5.2	56.2	85.9	41.5	49.0	57.0
PLLaVA-7B	38.0	35.8	36.3	39.9	38.0
Video-LLaVA	36.8	36.1	33.7	40.5	37.0
WorldAgent (Qwen3VL-8B)	**64.0**	62.0	41.8	47.0	53.4
WorldAgent (GPT-5.2)	57.3	**86.1**	**42.4**	**51.0**	**58.1**

consistently achieves strong performance across both benchmarks. When we use GPT-5.2 as the backbone, WorldAgent outperforms all baselines across all datasets and improves over its backbone by approximately 10% on Physion OCD and 6% on Physion v1.5 OCP. When we use the Qwen3VL-8B backbone, WorldAgent also demonstrates consistent gains and notably outperforms baselines with substantially larger model sizes (e.g., GPT-5.2 and Qwen3VL-30B-A3B) on most tasks. These results highlight the effectiveness of WorldAgent in both recognizing critical physical moments and predicting future physical dynamics. Moreover, we observe that the performance gains of WorldAgent scale with its underlying VLM backbone.

Results on PhysBench in Table 2 further support this conclusion. Across different backbone models, WorldAgent consistently improves average performance. Using Qwen3VL-8B as the backbone, WorldAgent improves performance by 2.5% on property-related questions and 2.1% on dynamics-related questions. When we use GPT-5.2 as the backbone, WorldAgent achieves gains of 1.1% on property questions and 2.0% on dynamics questions. The improvements on these two types of questions provide direct evidence of the effectiveness of the physical event memory and dynamics simulation tools. Improvements on relationship-based questions remain more limited, since most of these questions are grounded in static scenes (e.g., a single image) and do not benefit from WorldAgent's physical video reasoning.

4.3 Efficiency of WorldAgent

Table 3 shows that WorldAgent's main computational cost comes from physical event memory construction, which requires multiple short-clip captioning operations. We further compare WorldAgent with the agent-based VideoTree method on Physion. WorldAgent is $> 2\times$ faster on OCD and $> 3\times$ faster on OCP. This efficiency comes from its single-pass frame selection, unlike VideoTree's iterative video exploration. Also, VideoTree captions all video frames, whereas WorldAgent captions only selected frames, further reducing cost.

Table 3: Comparison of the runtime of WorldAgent and other baselines on the Physion dataset.

Method	Samp.	Mem.	Sim.	OCD	OCP
VideoTree	–	–	–	6.764	3.841
WorldAgent (GPT-5.2)	0.152	2.620	0.712	2.950	0.891

Table 4: Ablation study on Physion and Physion v1.5. Samp. = Event-centric sampling, Mem. = Physical Event Memory, Sim. = Dynamics Simulation.

Components			Physion		Physion v1.5	
Samp.	Mem.	Sim.	OCD	OCP	OCD	OCP
–	–	–	70.3	53.0	75.3	55.3
✓	–	–	70.3	53.4	76.9	55.4
–	✓	–	72.5	–	79.2	–
–	–	✓	–	58.0	–	59.4
✓	✓	–	74.1	–	79.6	–
✓	–	✓	–	58.8	–	62.4
✓	✓	✓	**74.1**	**58.8**	**79.6**	**62.4**

4.4 Ablation Study

We conduct all ablations using WorldAgent with the Qwen3VL-8B backbone.

4.4.1 Effectiveness of Physical Event Memory

As shown in Table 4, incorporating Physical Event Memory substantially improves performance on physical questions that require perceptual reasoning (OCD task), yielding a 2.2% improvement on Physion and a 3.9% improvement on Physion v1.5. These results indicate that converting physical video content into structured textual represen-

Table 5: Effect of consecutive frame comparison and summarization in Physical Event Memory on OCD.

Setting	Phy.	Phy. v1.5
w/o Cons.	72.5	77.4
w/o Sum.	69.0	77.5
Full	74.1	79.6

Table 6: Effect of the word length of the physical event memory and dynamics simulation on Physion.

Length	Mem.	Sim.
100	74.1	58.8
200	71.2	58.6
300	71.5	56.9

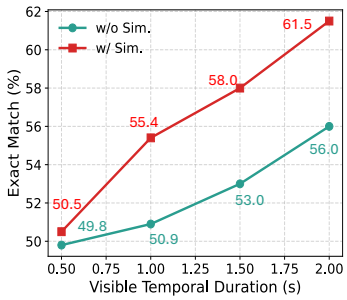

Figure 4: Effect of the dynamics simulation tool on Physion object prediction under varying visible durations.

tations plays a crucial role in enabling VLMs to perform perceptual reasoning in physical scenarios. Table 5 further demonstrates that both consecutive frame comparison and summarization contribute to the observed performance gains, which suggests that capturing both first-order state transitions and higher-order dynamics is essential for constructing high-quality physical event memory. Finally, the results in Table 6 show that increasing memory length can degrade performance, which indicates that the final memory should remain compact. Since the memory is intended to complement the visual inputs, excessively long memories may reduce attention to the visual evidence.

4.4.2 Effectiveness of Dynamics Simulation

The results in Table 2 show that incorporating dynamics simulation makes WorldAgent particularly effective at solving physical dynamics questions. The ablation study in Table 4 provides direct evidence that dynamics simulation substantially benefits the object contact prediction task, which requires reasoning about future physical state evolution. Complementary results in Fig. 4 confirm that the dynamics simulation module yields consistent performance gains across videos of varying lengths, from 0.5 to 2.0 seconds. Notably, longer visible video contexts lead to more accurate simulation outcomes. Together, these findings suggest that existing VLMs possess latent capabilities for reasoning about unseen and counterfactual physical events but require an explicit dynamics simulation step to effectively leverage them. Consistent with the observations for Physical Event Memory, Table 6 indicates that simulation outputs should remain concise to achieve optimal performance.

Table 7: Evaluating Qwen3VL-8B on the Physion benchmarks using event-centric frames sampled with different surprise scores generated by V-JEPA2 (ViT-G).

Frame Selection	Physion		Physion v1.5	
	OCD	OCP	OCD	OCP
Min. Surprise	61.5	52.1	66.9	53.0
Inter. Surprise	67.9	53.3	74.0	55.3
Max. Surprise	**70.3**	**53.4**	**76.9**	**55.4**

Table 8: Effect of different V-JEPA2 backbones to event-centric sampling on Physion and Physion v1.5 datasets.

V-JEPA2 Backbone	Physion		Physion v1.5	
	OCD	OCP	OCD	OCP
ViT-L	63.2	51.9	76.1	55.0
ViT-H	67.8	**53.6**	76.2	55.2
ViT-G	**70.3**	53.4	**76.9**	**55.4**

4.4.3 Effectiveness of Event-Centric Sampling

In Table 3, we show that event-centric frame sampling alone yields noticeable performance gains on both Physion datasets across all settings. Notably, event-centric sampling further amplifies the effectiveness of both Physical Event Memory and Dynamics Simulation. These results demonstrate that event-centric frames effectively capture salient physical events and provide a foundation for predicting counterfactual physical dynamics.

High-surprise frames yield the best performance. In Table 7 , we compare sampling strategies based on minimal, intermediate, and maximal surprise scores. The results reveal a clear positive correlation between model performance and surprise score across all tasks, with higher-surprise frames consistently achieving better accuracy.

Stronger V-JEPA2 improves frame sampling. In Table 8, we show that the expressive capacity of the V-JEPA2 correlates positively with physical reasoning performance. Stronger backbones (e.g., ViT-G) consistently outperform ViT-H and ViT-L and enable more effective sampling.

5 Conclusion

In this work, we propose WorldAgent, a training-free VLM agent for physical world understanding. WorldAgent performs explicit physical event-level analysis using two agentic tools, Physical Event Memory and Dynamics Simulation, and further enhances physical reasoning through event-centric sampling. WorldAgent achieves state-of-the-art performance on three physical world benchmarks that cover diverse scenarios and reasoning types.

Limitations

First, this work mainly evaluates WorldAgent on short physical video clips, providing limited evaluation of its long-horizon physical world understanding ability. Second, although we believe event-centric sampling could benefit commonsense video understanding tasks, we do not conduct experiments on general video QA in this work.

Ethical Considerations

All datasets, benchmarks, and baseline models used in this work are publicly available research artifacts, with the exception of GPT-5.2, which is accessed through a proprietary API. We cite the original creators of these artifacts and follow the corresponding licenses, terms of use, and distribution policies. This study does not involve collecting new human-subject data, personally identifiable information, or private user data. AI-assisted tools were used only for grammar checking, language polishing, and improving clarity of expression. They were not used to generate experimental results, conduct analysis, or make scientific claims.

References

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, and 1 others. 2025. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint*.

Tayfun Ates, M Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. 2022. Craft: A benchmark for causal reasoning about forces and interactions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2602–2627.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. Qwen3-vl technical report. *arXiv preprint*.

Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G. Krishnan. 2025. Physics context builders: A modular framework for physical reasoning in vision-language models.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. 2023. V-jepa: Latent video prediction for visual representation learning. *arXiv preprint*.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. 2013. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, and 1 others. 2021. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint*.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. 2025. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint*.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint*.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. 2025. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. *ICLR*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. *CVPR*.

Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. 2022. Pip: Physical interaction prediction via mental simulation with span selection. *ECCV*.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. *ECCV*.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. 2025. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint*.

Google. 2025. A new era of intelligence with gemini 3. Accessed: 2025.

Vincent Le Guen and Nicolas Thome. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. *CVPR*.

Lawrence Jang, Yinheng Li, Dan Zhao, Charles Ding, Justin Lin, Paul Pu Liang, Rogerio Bonatti, and Kazuhito Koishida. 2024. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. *arXiv preprint*.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. 2023. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *arXiv preprint*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *CVPR*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. *EMNLP*.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. 2025a. Longvideoagent: Multi-agent reasoning with long videos. *arXiv preprint*.

Ruyang Liu, Shangkun Sun, Haoran Tang, Wei Gao, and Ge Li. 2025b. Flow4agent: Long-form video understanding via motion prior from optical flow. *ICCV*.

Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, and 1 others. 2025. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly. *arXiv preprint*.

Michael McCloskey. 1983. Intuitive physics. *Scientific american*.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *ICML*.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. 2020. Intphys: A framework and benchmark for visual intuitive physics reasoning.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, and 1 others. 2022. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. *CoRL*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint*.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. 2023. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *NeurIPS*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025b. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. *CVPR*.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint*.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025. Vca: Video curious agent for long video understanding. In *ICCV*.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. 2025. Generative frame sampler for long video understanding. *arXiv preprint*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. *ICLR*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint*.

Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B Tenenbaum, and Chuang Gan. 2024. Contphy: Continuum physical concept learning and reasoning from videos. *arXiv preprint*.

Appendix

A Implementation Details

All general-purpose VLM baselines uniformly sample 8 frames per video, following the evaluation protocol of PhysBench (Chow et al., 2025). VideoTree selects a variable number of frames across videos due to its adaptive sampling strategy. For video VLMs, the default frame configurations are adopted. WorldAgent performs event-centric sampling using the V-JEPA2 ViT-G checkpoint: it first uniformly samples 64 frames from each video and then selects 8 salient frames based on their surprise scores.

B Prompts

B.1 Prompts for Reasoning Type Categorization

The prompt on the right illustrates how the VLM manager categorizes the reasoning type of a physical query. The model outputs a reasoning label from perceptual, counterfactual, or predictive.

Reasoning Type Classification

You are a classifier for physical reasoning type.
Task: Given ONE question string, classify the reasoning type required to answer the question.
Choose **EXACTLY ONE** of the following three classes:

- **perceptual**: The question can be answered by observing the visible content, including current or past states in the video or image. No future prediction or hypothetical intervention is required.
- **predictive**: The question requires predicting a future physical state or outcome beyond the observed frames (e.g., what will happen next).
- **counterfactual**: The question requires reasoning about a hypothetical or altered scenario that differs from what is observed (e.g., "if X were removed" or "if the object were heavier").

Output format:

```
{
  "reasoning_type": "perceptual" |
  "predictive" | "counterfactual"
}
```

Do not include any additional text, explanations, or keys.

B.2 Prompts for Physical Event Memory Construction

Below illustrates two prompts used for consecutive frame comparison and summarization in building physical event memory.

Prompt for Consecutive Frame Comparison

You are given two consecutive frames from a physical-world video: **Frame t** and **Frame t+1**.
Analyze the change between the two frames and explain the underlying cause, and output **only** a valid JSON object with the following fields:

```
{
  "physical_state_transition": "
      One sentence description of
      the observed change between
      Frame t and Frame t+1,
      expressed explicitly as a
      physical state transition",
  "underlying_physical_cause": "
      One sentence intuitive
      explanation of why this
      transition occurred, grounded
       in basic physical principles
       such as gravity, contact
      forces, inertia, friction,
      elasticity, or loss of
      support."
}
```

Rules:

- The `physical_state_transition` must fully represent the observed difference; do not describe the change elsewhere.
- Output JSON only.
- Do not include any extra text outside the JSON.
- Base the explanation strictly on visual evidence and common-sense physical intuition.

Prompt for Memory Summarization

You are given sampled frames from a physical world video and comparisons between consecutive frames.

Frame 1-2: Difference: {} Explanation: {}
Frame 2-3: Difference: {} Explanation: {}
Frame 3-4: Difference: {} Explanation: {}
...

Using the frame comparisons and by re-examining the video to capture higher-order dynamics, briefly summarize the key physical events that occur in the video.

Keep the description under 100 words.

Do not include any additional text, explanations, or keys.

B.3 Prompts for Dynamics Simulation

Below illustrates the **simplified** prompt used to generate a simulation outcome given an input query and sampled physical video frames.

Prompt for Dynamics Simulation

Using the events observed so far and the context mentioned in the question, accurately simulating the scene's likely evolution in the given context using physical knowledge and object-level physical properties (e.g., brittleness, stiffness, softness, smoothness, malleability, sharpness, elasticity).

The question is: <question>

Summarize the expected simulation in a single short paragraph (maximum 100 words).

B.4 Prompts for Benchmark Evaluation

In this section, we illustrate the evaluation prompts used for the Physion, Physion v1.5, and PhysBench benchmarks. Physion and Physion v1.5 use an identical evaluation prompt.

Evaluation Prompt for Physion and Physion v1.5

You are given a physical simulation video and a question.

Q: Is the red object going to contact the yellow area?

CRITICAL OUTPUT INSTRUCTION: Your response must be (A or B) only.

Please provide your analysis in the required format. Do not include any other text, explanation, context, preamble, or punctuation before or after the answer.

Evaluation Prompt for PhysBench

This is a series of images sampled from the beginning to the end of a video, based on the series of images, output the best option for the question.

Q: <question>

CRITICAL OUTPUT INSTRUCTION: Your response must be (A or B or C or D) only.

Please provide your analysis in the required format. Do not include any other text, explanation, context, preamble, or punctuation before or after the answer.

C Additional Experiments

C.1 Physion Datasets Question Design

As the Physion datasets were not originally intended for question answering, we adopt the question prompt that yields the best performance with Qwen3VL-8B on the Physion OCD task as the evaluation prompt for all baselines.

Table 9: Effect of prompt phrasing on Physion OCD task on Qwen3-VL-8B.

Prompt	EM@1
Will the red object hit the yellow object?	62.58
Will the red object contact the yellow object?	66.92
Does red object contact the yellow object?	61.58
Is red object going to hit the yellow area?	68.42
Is red object going to contact the yellow area?	70.30

Table 10: Effect of event-centric sampling across different VLMs on Physion and Physion v1.5 datasets.

VLM Backbone	Samp.	Physion OCD	Physion OCP	Physion v1.5 OCD	Physion v1.5 OCP
InternVL3.5-8B	–	71.5	55.8	70.1	55.2
	✓	72.2	55.2	72.3	55.4
InternVL3.5-30B-A3B	–	70.8	54.9	72.5	53.9
	✓	70.9	55.3	72.5	54.5
Qwen3VL-30B-A3B	–	71.4	54.4	74.8	56.4
	✓	71.0	55.3	74.9	57.2

C.2 More Ablations

Event-centric sampling is backbone-agnostic. Table 10 shows consistent gains across diverse VLM backbones, including InternVL and Qwen variants, indicating that event-centric sampling generalizes well across existing VLMs.

Effect of Frame Number Table 11 reports the performance comparison between WorldAgent and Qwen3-VL-8B on the Physion datasets under varying numbers of input frames. The results show that increasing the number of frames yields limited performance gains for the baseline, while WorldAgent consistently outperforms the baseline across all settings.

Table 11: Effect of Frame Number.

Method	Frame #	v1.0 OCD	v1.0 OCP	v1.5 OCD	v1.5 OCP
Qwen3VL-8B	8	70.3	53.0	75.3	55.3
↳ WorldAgent	8	74.1	58.8	79.6	62.4
Qwen3VL-8B	16	68.2	53.3	75.0	56.1
↳ WorldAgent	16	73.2	58.8	79.9	63.1

Memory for Dynamic Simulation. We additionally evaluate whether physical event memory benefits the dynamics simulation tool. Table 12 shows that memory provides only marginal improvements on Physion OCP task and negatively impacts performance on Physion v1.5.

Table 12: Effect of using physical event memory for dynamics simulation.

Setting	Physion OCP	Physion v1.5 OCP
w/o Mem.	57.75	62.42
w/ Mem.	57.83	62.22