

# PHYSNOTE: SELF-KNOWLEDGE NOTES FOR EVOLVABLE PHYSICAL REASONING IN VISION-LANGUAGE MODEL

Sinin Zhang<sup>1</sup>, Yunfei Xie<sup>2</sup>, Yuxuan Cheng<sup>1†</sup>, Haoyu Zhang<sup>3</sup>, Tong Zhang<sup>4</sup>

<sup>1</sup>The Chinese University of Hong Kong, Shenzhen, <sup>2</sup>Rice University,

<sup>3</sup>City University of Hong Kong, <sup>4</sup>Fudan University

## ABSTRACT

Vision-Language Models (VLMs) have demonstrated strong performance on textbook-style physics problems, yet they frequently fail when confronted with dynamic real-world scenarios that require temporal consistency and causal reasoning across frames. We identify two fundamental challenges underlying these failures: (1) spatio-temporal identity drift, where objects lose their physical identity across successive frames and break causal chains, and (2) volatility of inference-time insights, where a model may occasionally produce correct physical reasoning but never consolidates it for future reuse. To address these challenges, we propose PhysNote, an agentic framework that enables VLMs to externalize and refine physical knowledge through self-generated “Knowledge Notes.” PhysNote stabilizes dynamic perception through spatio-temporal canonicalization, organizes self-generated insights into a hierarchical knowledge repository, and drives an iterative reasoning loop that grounds hypotheses in visual evidence before consolidating verified knowledge. Experiments on PhysBench demonstrate that PhysNote achieves 56.68% overall accuracy, a 4.96% improvement over the best multi-agent baseline, with consistent gains across all four physical reasoning domains.

## 1 INTRODUCTION

Building autonomous agents that can act safely in the physical world requires *physical reasoning*: the ability to connect observations with the causal mechanisms that generate motion, contact, and state changes over time. Humans acquire this ability early as intuitive physics and apply it effortlessly, yet replicating it in artificial systems remains an open challenge. This capability links visual perception to prediction and control, and it is therefore central to reliable embodied interaction, from robotic manipulation to autonomous navigation.

Recent Vision-Language Models (VLMs) have shown strong performance on static scene understanding and textbook-style physics problems He et al. (2024); Yue et al. (2024). With increased pre-training scale, models Achiam et al. (2023); Guo et al. (2025); Comanici et al. (2025) can often infer mathematical solutions from single-frame schematics and recognize fundamental physical concepts. However, much of this progress reflects competence in “text-based physics” Qiu et al. (2025); Zhang et al. (2025), where the task is presented as a symbolic or linguistic problem and the supervision signal is contained in the prompt. In such settings, correct answers can be produced without constructing a temporally grounded representation of the underlying physical process Chow et al.; Yang et al. (2025). This limitation becomes clear when the input shifts from static images to multi-frame sequences that contain temporal evolution and causal interaction among objects LeCun (2022). In these dynamic settings, the model must track objects across time, detect contact events, and maintain consistent causal chains, capabilities that static, text-based tasks never require.

Even state-of-the-art VLMs often fail at these requirements. An extensive evaluation of 75 VLMs on PhysBench Chow et al. shows that most models achieve only around 40% accuracy on physical reasoning tasks, far below human performance, with perceptual errors and knowledge gaps accounting for the majority of failures. Notably, this deficit does not diminish with larger model size, more training data, or additional input frames, suggesting that physical understanding is not an emergent

property of scale alone. Objects frequently lose their physical identity across successive frames, and causal chains can be broken by hallucinatory transitions. These findings indicate that current VLMs lack mechanisms to reliably anchor perception and reasoning in both time and space, which limits their applicability to embodied tasks that demand precise physical inference, such as manipulation where misidentifying an object or misjudging a trajectory can cause operational failures Chow et al..

We identify two fundamental challenges that underlie these failures in dynamic physical reasoning: (1) **Spatio-Temporal Identity Drift in Perception** and (2) **Volatility of Inference-time Insights**.

For the first challenge, dynamic physical reasoning requires object permanence and traceability of states across time. However, standard VLMs frequently exhibit “identity drift,” in which the semantic representation of an object shifts or disappears across successive frames. Without an explicit mechanism for spatio-temporal anchoring, a causal chain, such as a collision that leads to a particular trajectory, becomes fragmented, which prevents the model from forming a coherent representation of scene dynamics.

For the second challenge, human physicists do not treat every new problem as a blank slate; they externalize insights, refine heuristics, and consolidate knowledge into reusable notes. In contrast, existing reasoning paradigms in VLMs are largely *volatile*. While a model may occasionally generate profound physical insights during a specific reasoning trace, these insights are transient and are not consolidated into a persistent knowledge store. Consequently, the model remains a “goldfish-memory physicist,” failing to evolve its internal world model through experience or self-reflection.

To address these challenges, we introduce **PhysNote**, an agentic framework that enables VLMs to externalize and refine physical knowledge through self-generated “**Knowledge Notes**.” PhysNote is designed to support dynamic reasoning by iterating over observation of a sequence, retrieval of relevant prior notes, and synthesis of updated insights that remain available for future problems.

PhysNote comprises three core components. To mitigate Spatio-Temporal Identity Drift in Perception, PhysNote employs a **Spatio-Temporal Canonicalization** module (Sec. 3.1) that assigns immutable numeric identifiers to visual tokens and enforces a *Triadic Observation Template*, providing stable visual anchors for object permanence across temporal sequences. To mitigate Volatility of Inference-time Insights, PhysNote introduces a **Hierarchical Knowledge Architecture** (Sec. 3.2) that externalizes self-generated insights into a three-tier repository spanning domain-level strategies and task-specific procedural details. This repository is updated through autonomous pruning and reflection, which preserves high-fidelity knowledge while systematically discarding erroneous biases. Connecting these two components, an **Agentic Reasoning** engine (Sec. 3.3) drives an iterative Hypothesis-Evidence-Validation loop: the agent gathers grounded visual evidence through the canonicalized anchors, consults the Knowledge Notes for relevant physical heuristics, and consolidates verified insights back into the repository after each problem.

Experiments on PhysBench Chow et al. show consistent gains for open-source models. On the large-scale test set, PhysNote achieves 56.68% overall accuracy, a 4.96% absolute improvement over the best multi-agent baseline, with balanced gains across all four physical domains. On the validation set, PhysNote reaches 72.86%, exceeding the Qwen2.5-VL-72B baseline by 3.01%. Ablation studies confirm that the synergy between the InfoAgent and Knowledge Notes is essential: neither component alone fully accounts for the gain. Qualitative analysis further shows that PhysNote prevents identity drift in dynamic scenes and enables the model to detect fine-grained physical cues that standard VLMs overlook.

## 2 RELATED WORK

### 2.1 PHYSICAL REASONING MODELS

The quest for physical intelligence in AI has evolved from symbolic engines to data-driven neural architectures. Early works primarily focused on Physics-Specialized Models Guen & Thome (2020); Duan et al. (2022), which utilize neural networks to approximate differential equations or Lagrangian dynamics. While precise, these models are often task-specific (e.g., fluid dynamics or rigid-body collisions) and require structured representations that are unavailable in raw visual inputs.

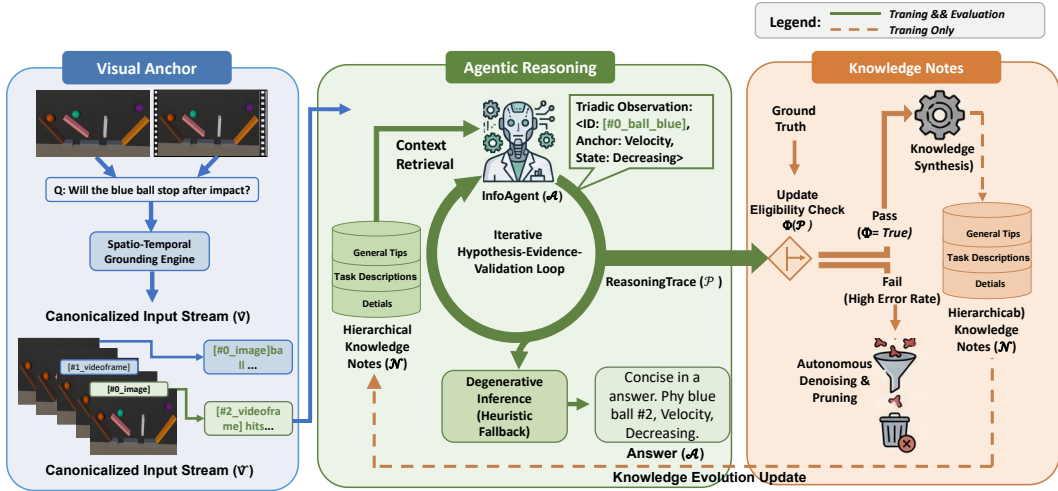


Figure 1: Overview of the PhysNote framework, which operates across three interconnected spaces to enable evolvable physical reasoning. **Visual Anchors** (left): Given a question  $Q$  and visual inputs  $V$  (images or video), a Spatio-Temporal Grounding Engine assigns each visual entity an immutable identifier (e.g., `[#0_ball_blue]`) to produce a canonicalized input stream  $\hat{V}$ , establishing object permanence across frames. **Agentic Reasoning** (center): The InfoAgent  $\mathcal{A}$  retrieves task-relevant context from Hierarchical Knowledge Notes  $\mathcal{N}$ , organized into General Tips, Task Descriptions, and Details. It then executes an iterative Hypothesis-Evidence-Validation loop, extracting structured Triadic Observations  $\langle \text{ID}, \text{Anchor}, \text{State} \rangle$  from the visual data to build a grounded reasoning trace  $\mathcal{P}$ . If the evidence gap remains unclosed, a Degenerative Inference mode applies heuristic fallback to produce the final answer  $\mathcal{A}$ . **Knowledge Notes** (right, training only): An Update Eligibility Check  $\Phi(\mathcal{P})$  evaluates whether the reasoning trace meets the criteria for knowledge consolidation. Eligible traces ( $\Phi = \text{True}$ ) are synthesized into updated Knowledge Notes; task nodes with persistently high error rates are removed through Autonomous Denoising and Pruning, ensuring the knowledge base evolves toward higher fidelity.

Recently, the focus has shifted to Vision-Language Models (VLMs) as generalist observers Xu et al. (2024). However, these models often suffer from Spatio-Temporal Identity Drift Yang et al. (2025). VLMs lack spatial reasoning and temporal grounding, causing objects to lose their physical identity across sequences. In contrast, PhysNote addresses this by replacing implicit visual attention with a deterministic Spatio-Temporal Canonicalization module, ensuring that visual anchors remain consistent throughout the reasoning trajectory.

## 2.2 AGENT-BASED PHYSICAL UNDERSTANDING

The integration of Vision-Language Models (VLMs) into Autonomous Agents has enabled models to tackle complex physical tasks via iterative prompting, such as Chain-of-Thought Wei et al. (2022) or ReAct paradigms Yao et al. (2022). To enhance physical grounding, specialized frameworks like PhysAgent Chow et al. incorporate a “reason-act-observe” loop that interfaces with external visual tools (e.g., SAM Kirillov et al. (2023), Depth Anything Yang et al. (2024)) to extract precise geometric and state information. Alternatively, Physics Context Builders (PCBs) Balazadeh et al. (2025) employ modular, fine-tuned smaller VLMs to generate detailed scene descriptions that serve as auxiliary reasoning context. Despite their successes, these approaches either rely on ephemeral reasoning traces that dissipate once the context window closes, or require static fine-tuning that lacks the capacity for autonomous knowledge evolution.

In contrast, PhysNote enables the persistent externalization and autonomous refinement of physical heuristics across a broader spectrum of dynamic scenarios without the need for expensive parameter fine-tuning or specialized external tools.

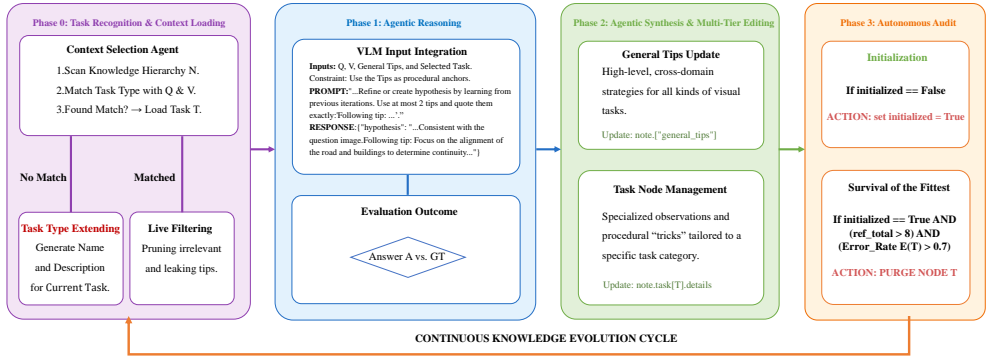


Figure 2: The Knowledge Note pipeline: usage and evolution. During inference, the system retrieves task-relevant notes from the Hierarchical Knowledge Notes  $\mathcal{N}$  based on the canonicalized inputs  $(\hat{Q}, \hat{V})$ , and feeds them into the InfoAgent’s iterative Hypothesis-Evidence-Validation loop to produce a grounded answer. During training, the resulting reasoning trace  $\mathcal{P}$  undergoes an Update Eligibility Check  $\Phi(\mathcal{P})$ : eligible traces trigger knowledge synthesis to update the notes, while task nodes with persistently high error rates are pruned, enabling the knowledge base to evolve autonomously toward higher fidelity over successive training batches.

### 3 METHODOLOGY

Current Vision-Language Models (VLMs) can analyze single-frame attributes and solve text-based physics problems, yet they frequently fail in multi-frame dynamic scenarios that demand temporal consistency and causal modeling. This failure stems not from a lack of internal physical knowledge, but from an inability to connect abstract principles to raw, noisy visual signals, much like a skilled mechanic who still requires a *checklist* to inspect the right components in the right order.

To bridge this gap, we propose *PhysNote*, a multi-agent framework that enables a VLM to generate, consult, and refine its own “Knowledge Notes.” As illustrated in Figure 1, the framework operates through a structured pipeline: it first organizes raw visual data by assigning unique identifiers, then identifies the task type and retrieves relevant notes, executes an iterative reasoning loop, and finally reflects on the outcome to update the notes. This pipeline comprises three core stages: (1) *Spatio-Temporal Canonicalization* (Section 3.1), which creates stable visual anchors; (2) *Hierarchical Knowledge Architecture* (Section 3.2), which stores task-specific insights; and (3) *Agentic Reasoning* (Section 3.3), which performs information discovery and knowledge consolidation.

#### 3.1 VISUAL ANCHORS: SPATIO-TEMPORAL CANONICALIZATION

A fundamental prerequisite for physical reasoning is object permanence: the ability to track an entity across time. Conventional VLMs often struggle with “identity drift,” where an object is recognized differently from one frame to the next. To mitigate this, given a raw question  $Q$  and its associated visual inputs  $V$  (images or video frames), we implement a canonicalization pipeline that transforms  $(Q, V)$  into a grounded reference pair  $(\hat{Q}, \hat{V})$ .

**Spatio-Temporal Indexing.** For any input  $V$ , the system discriminates between static images and video sequences to assign immutable identifiers, such as `[#0image]` or `[#1videoframe3]`. These identifiers are embedded into the original question text to produce  $\hat{Q}$ , so that every reference to a visual element is explicit and unambiguous. The visual inputs are likewise repackaged with their assigned identifiers to produce  $\hat{V}$ . Together, the tagged visuals  $\hat{V}$  and the annotated question  $\hat{Q}$  establish a consistent referencing scheme for all downstream processing.

**Triadic Observation Template.** While Spatio-Temporal Indexing resolves temporal ambiguity by tagging each frame, an additional mechanism is needed to resolve entity ambiguity within each frame. To bridge the gap between raw pixels and symbolic reasoning, we require every observation to answer three questions: *which frame* it comes from, *which object* it describes, and *what property*

is observed. Every observation  $O$  extracted by the InfoAgent (Section 3.3) from  $\hat{V}$  is accordingly formalized as:

$$O = \left( \underbrace{v}_{\text{Visual ID}}, \underbrace{e}_{\text{Visual Anchor}}, \underbrace{d}_{\text{Description}} \right) \quad (1)$$

Here  $v$  is the identifier assigned during Spatio-Temporal Indexing,  $e$  denotes the physical entity of interest (e.g., a specific ball or surface), and  $d$  states a single observable property of that entity (e.g., position, velocity, or contact state).

By enforcing this structured format, every observation is traceable to a specific frame and entity, preventing information loss when observations are passed between processing stages (Section 3.3). This protocol compels the model to decompose complex dynamic scenes into a series of verifiable, grounded facts rather than vague or misidentified references.

### 3.2 KNOWLEDGE HIERARCHY AND TASK-ORIENTED SELECTION

To store and retrieve physical insights, we externalize self-generated knowledge into a three-tier hierarchy  $\mathcal{N}$ , termed *Knowledge Notes*. This repository is structured to balance universal physical principles with task-specific heuristics, comprising three levels. (1) **General Tips** capture high-level, cross-domain strategies that apply broadly across different physical scenarios. (2) **Task Descriptions** provide natural-language definitions that delineate the boundaries of a specific physical phenomenon, specifying both textual and visual cues that characterize the task category. (3) **Task Details** record specialized observations and procedural guidance tailored to each task category, accumulated through the reflective process described in Section 3.3.

The selection process is governed by a *Context Selection Agent*. Given the canonicalized inputs  $(\hat{Q}, \hat{V})$ , the agent evaluates them against  $\mathcal{N}$  to identify the most relevant task node  $T$ . Unlike traditional retrieval, this agent operates with *Global Visibility*, meaning it can synthesize information across the entire hierarchy to determine if the current scenario matches an existing node or represents a novel physical task. If the system fails to find a suitable match and the expansion mode is active, it invokes a *Note Discovery* routine to define a new task node  $T$ . By observing the global state of  $\mathcal{N}$ , the system avoids redundant definitions, ensuring that categories such as “Free Fall” and “Projectile Motion” remain semantically distinct yet functionally related.

To ensure the long-term reliability of  $\mathcal{N}$  and prevent “knowledge pollution” from erroneous reasoning traces, we implement an autonomous pruning mechanism. The core idea is simple: every time a knowledge node is consulted, the system records whether the inference succeeded or failed. For each task node  $T$ , let  $n_+$  and  $n_-$  denote the cumulative counts of correct and incorrect outcomes. The empirical error rate is:

$$E(T) = \frac{n_-}{n_+ + n_-} \quad (2)$$

Once  $E(T)$  exceeds a predefined threshold  $\tau$  after a minimum of  $n_{\min}$  activations, the node  $T$  and its associated *Task Details* are purged from  $\mathcal{N}$ . This “survival of the fittest” logic ensures that the knowledge base evolves towards higher precision, systematically discarding speculative or biased information that may have been generated during earlier iterations.

### 3.3 AGENTIC REASONING: THE INFOAGENT LOOP AND REFLECTION

The core reasoning engine, the *InfoAgent*, is designed to transform high-noise visual environments into structured, textbook-style physics problems. This is achieved through an iterative *Hypothesis–Evidence–Validation* loop, which is capped at a maximum of three iterations ( $i \in \{1, 2, 3\}$ ) to maintain computational efficiency and prevent the agent from accumulating contradictory hypotheses.

In each iteration  $i$ , the agent assumes three functional roles sequentially. First, based on the canonicalized inputs  $(\hat{Q}, \hat{V})$ , the retrieved Knowledge Notes, and any observations from previous iterations, it generates a candidate physical explanation  $H_i$  together with a preliminary candidate answer  $A_{\text{cand}}$ , accompanied by a set of actionable queries  $Queries_i$  encoded as `<info></info>` and `<attention></attention>` signals. Second, these queries are executed against the canonicalized visual set  $\hat{V}$  to extract micro-facts  $Facts_i$ . To ensure objective grounding, each fact must strictly adhere to the *Triadic Observation Template* defined in Section 3.1. Third, a specialized

validator module assesses whether  $Facts_i$  provides sufficient evidence to support  $A_{\text{cand}}$ . If critical information is missing, the validator provides explicit negative feedback, marking specific pieces of evidence as “unobtainable” to refine the hypothesis in iteration  $i + 1$ .

In scenarios where the information gap remains unclosed after the iteration limit, the system enters a *Degenerative Inference Mode*. In this state, the agent supplements grounded evidence with its internal world knowledge (i.e., heuristic assumptions) to produce a final answer  $A$ . While this fallback ensures that every question receives an answer, the resulting reasoning is less trustworthy because it relies on unverified priors.

**Update Eligibility and Reflection.** After the InfoAgent produces an answer, the system determines whether the resulting reasoning trace should be used to update the Knowledge Notes  $\mathcal{N}$ . We define a reasoning trace  $P$  as the full record of hypotheses, queries, and observations across all iterations. During the training phase, where ground-truth labels are available,  $P$  is eligible for knowledge consolidation only when it represents a fully grounded success, meaning four conditions must hold simultaneously: (i) the final answer matches the ground truth, (ii) the Degenerative Inference Mode was not activated, (iii) no heuristic assumption was used in place of grounded evidence, and (iv) the validated micro-facts contain explicit physical operators (e.g., *trajectory*, *collision*, *contact*) rather than purely linguistic descriptors. Denoting these four predicates as  $c$ ,  $f$ ,  $a$ , and  $d$  respectively, eligibility is expressed as their conjunction:

$$\Phi(P) = c(P) \wedge \neg f(P) \wedge \neg a(P) \wedge d(P) \quad (3)$$

This strict gate prevents the system from consolidating unverified biases into its long-term knowledge. The reflection phase concludes with a dual-path update logic. If  $P$  is both correct and decisive, the system performs a reflective update to the *Task Details* and *General Tips* within  $\mathcal{N}$ , extracting the underlying physical principles from the trace. Conversely, if the model fails but a ground truth exists, a *Tip Discovery* routine is initiated to propose 1–2 actionable, observable tips that are appended to the corresponding *Task Details* for future reuse. This reflective cycle ensures that the system learns not only from its grounded successes but also from its perceptual and logical failures.

## 4 EXPERIMENTS

### 4.1 IMPLEMENTATION DETAILS

For implementation, we employ Qwen2.5-VL-72B-Instruct Bai et al. (2025) as the base. Video inputs are processed by sampling 4 frames uniformly across the sequence at a resolution of 512 pixels. Knowledge management leverages a hierarchical JSON structure and all-MiniLM-L6-v2 embeddings for task retrieval. To maintain knowledge purity, the pruning mechanism deprecates any node  $T$  with an error rate  $E(T) > 0.7$  after at least 8 activations. The *InfoAgent* reasoning is capped at 3 iterations, utilizing a 15-primitive keyword filter to gate the Update Eligibility criterion  $\Phi(P)$ .

### 4.2 DATASET AND BASELINE

#### 4.2.1 DATASET SETUP

We evaluate PhysNote on PhysBench Chow et al., a large-scale interleaved video-image-text dataset. For the knowledge evolution phase, we sample 500 entries from the official training set. Main quantitative results are reported on the full test set (10,000 entries), while ablation studies are conducted on the validation set (200 entries). All subsets maintain an equal distribution across the four major physical domains: (S1) **Physics-based dynamics** (e.g., collisions and projectile motion), (S2) **Physical object relationships** (e.g., spatial layout and support relations), (S3) **Physical scene understanding** (e.g., lighting and environment context), and (S4) **Physical object properties** (e.g., mass, friction, and elasticity).

#### 4.2.2 BASELINES

We compare *PhysNote* against a diverse set of baselines categorized into three groups: (1) **Open-source VLMs**, including LLaVA-series Liu et al. (2023), InternVL1.5 Chen et al. (2024), and

Mantis Jiang et al. (2024); (2) **Closed-source VLMs**, such as GPT-4o OpenAI (2024), Gemini-1.5 Reid et al. (2024), and Claude-3.5 Anthropic (2024); and (3) **Multi-agent systems**, specifically PhysAgent Chow et al., to evaluate the effectiveness of our knowledge-evolution framework. To establish an upper bound for physical reasoning, we also include human performance as a reference. All models are evaluated under a zero-shot setting to ensure a fair comparison of their inherent physical understanding and reasoning capabilities.

### 4.3 QUANTITATIVE ANALYSIS

#### 4.3.1 MAIN RESULTS

Table 1 presents the performance of various models on the PhysBench dataset. Our proposed *PhysNote* consistently outperforms all baseline models across all four physical domains (S1–S4). Specifically, *PhysNote* achieves an average score of 56.68%, which is a 4.96% absolute improvement over the state-of-the-art multi-agent baseline, *PhysAgent*.

Compared to individual open-source and closed-source VLMs, the advantage of our framework is even more significant. While closed-source models like GPT-4o and Gemini-1.5-Pro show competitive results in specific categories, they often struggle with temporal consistency in dynamic scenes (S1) and complex object relationships (S2). In contrast, *PhysNote* shows a balanced performance gain, particularly in Physics-based dynamics (S1) and Physical object relationships (S2). These results suggest that externalizing physical knowledge into structured notes effectively helps the model bridge the gap between raw visual inputs and high-level physical reasoning.

Table 1: **Quantitative results on PhysBench.** Accuracy comparison across open-source, closed-source, and multi-agent systems on four physical reasoning categories (S1–S4). **Bold** and underline indicate the best and second-best results within each category.

Category	Model	S1	S2	S3	S4	Avg.
<b>Baselines</b>	Human	97.10	95.67	94.91	95.68	95.87
<b>Open-source</b>	LLaVA-1.5-13B	41.31	42.50	34.40	44.38	40.45
	LLaVA-1.6-vicuna-7B	40.26	<u>59.72</u>	<b>38.60</b>	42.65	42.28
	InternVL-Chat-1.5	<b>53.08</b>	<b>70.14</b>	37.01	44.78	<b>47.51</b>
	PLLaVA-13B	39.91	38.33	31.52	40.76	37.70
	Mantis-siglip-llama3	42.47	32.78	<u>36.83</u>	37.51	37.64
	LLaVA-interleave-dpo	47.97	42.67	<u>33.73</u>	38.78	40.83
<b>Closed-source</b>	GPT-4o	56.91	<b>64.80</b>	30.15	<b>46.99</b>	<b>49.49</b>
	Gemini-1.5-pro	<u>57.26</u>	<u>63.61</u>	<b>36.52</b>	41.56	<u>49.11</u>
	Claude-3.5-sonnet	46.46	<u>41.11</u>	27.89	37.60	<u>38.05</u>
	Gemini-1.5-flash	<b>57.41</b>	52.24	34.32	40.93	46.07
<b>Multi-agent</b>	PhysAgent	58.20	65.40	<u>38.10</u>	45.20	51.72
	<b>PhysNote (Ours)</b>	<b>62.45</b>	<b>72.10</b>	<b>42.33</b>	<b>49.85</b>	<b>56.68</b>

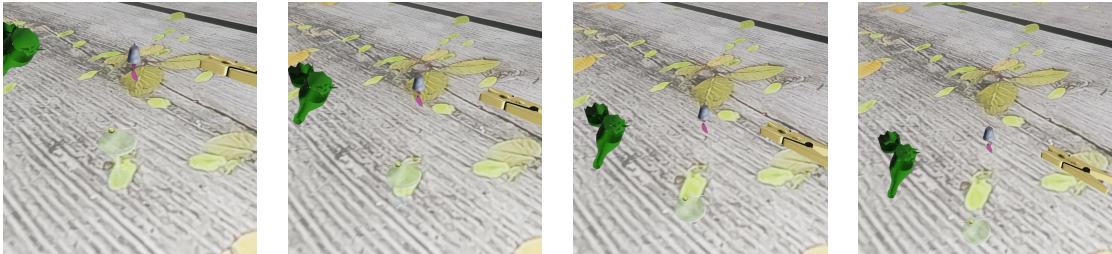
#### 4.3.2 ABLATION RESULTS

Table 2: **Ablation study on PhysBench validation set.** Using Qwen2.5-VL-72B-Instruct as the base model, we evaluate the contribution of each component across four task categories. **Bold** indicates the best result; underline indicates the second best.

Model Configuration	Dynamics	Scene	Relationships	Property	Overall Avg.
Baseline (VLM only)	<u>70.89</u>	54.05	<u>73.91</u>	<b>78.38</b>	<u>69.85</u>
Baseline+InfoAgent	63.29	54.05	71.74	67.57	64.32
Baseline+Note	60.76	<u>56.76</u>	<b>78.26</b>	<u>75.68</u>	66.83
<b>Ours (Full)</b>	<b>74.68</b>	<b>62.16</b>	<b>78.26</b>	72.97	<b>72.86</b>

Table 2 shows the contribution of each component to the overall performance. We observe that simply adding the *InfoAgent* or *Knowledge Notes* alone does not always lead to immediate improvements. In fact, the “Baseline+InfoAgent” configuration shows a slight performance drop, likely because the agent acts as a stochastic parrot when it lacks structured guidance, relying on semantic correlations rather than true physical grounding.

However, when these components are integrated into our Full Model, we achieve the best performance in most categories, with an overall accuracy of 72.86%. This demonstrates that the synergy between agentic exploration and structured knowledge management is essential for solving complex physical tasks.



**Question:** Given the observed phenomena in the video, which of the following camera adjustments could create the depicted effect? A. Move parallel to the line between the broken wine bottle and the martini; B. The camera is farther away from the objects; C. Move parallel to the line between the peg and the martini; D. The camera moves upward or downward.

**Ground Truth: D**

Baseline	Ours
<p><b>Raw Reasoning:</b> Analyze changes in video frames: 1. <b>Objects’ Positions:</b> Relative positions appear consistent. 2. <b>Perspective Changes:</b> No significant change in angle, suggests camera is not moving closer/farther. 3. <b>Vertical Alignment:</b> Objects maintain alignment, indicating camera is not moving upward/downward.</p> <p><i>Observations:</i> - Opt A: Parallel move wouldn’t significantly alter scene. - Opt B: Farther away would show size change (not observed). - Opt C: Parallel move between peg and martini seems plausible. - Opt D: Moving upward/downward would change vertical alignment (not observed).</p>	<p><b>Raw Reasoning:</b> Tips applied: Consistent relative size observed between bottle/martini; stable camera distance. Peg/martini alignment stable; no parallel movement.</p> <p><b>Iteration 1-2:</b> Hypothesis: bottle/martini maintain size; peg/martini show no change in alignment. Gathered Truths: Peg’s tip aligns consistently with top rim. Fluid surface below rim appears constant.</p> <p><b>Iteration 3:</b> Comparison of relative sizes shows consistency (rules out Opt B). Peg/martini alignment constant (rules out Opt C). Gray bell-shaped object distance to edge decreased.</p> <p><b>Assumption Rationale:</b> Consistent alignment and stable sizes suggest unchanged distance. Gray object position supports the inference for vertical displacement.</p>
<b>Extract Answer: C</b>	<b>Extract Answer: D</b>

Figure 3: **Qualitative comparison on camera motion reasoning.** Given a video sequence and a multiple-choice question about camera adjustments, the baseline (left) fails to detect the vertical shift and relies on perceived stability to select a lateral motion (C). Our method (right) iteratively verifies geometric constraints, including object alignment and relative scale, across successive frames. By ruling out lateral and depth-based hypotheses through grounded evidence, it correctly identifies vertical camera displacement (D).

#### 4.4 QUALITATIVE ANALYSIS

To further demonstrate the reasoning process, we present two representative cases comparing *PhysNote* with the baseline.

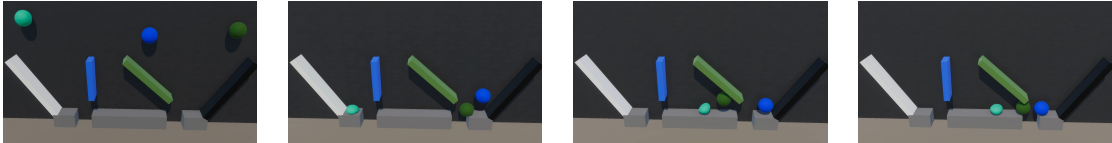
##### 4.4.1 CASE I: EVIDENCE-BASED VISUAL ANCHORING

As illustrated in Figure 3, the task requires identifying vertical camera displacement within a sequence of video frames. The Baseline model demonstrates a failure to maintain precise spatial tracking across the temporal sequence, leading to a misperception of stable vertical alignment. This perceptual drift causes the model to overlook subtle geometric shifts and instead rely on a plausible but physically incorrect lateral motion (Option C). In contrast, *PhysNote* utilizes its iterative

reasoning loop to systematically evaluate specific geometric constraints. By generating and verifying "Gathered Truths" regarding the relative size of objects and the alignment of the peg's tip, the framework rules out hypotheses involving depth and parallel movement. Specifically, by detecting the displacement of the gray object relative to the frame boundary, PhysNote correctly identifies the vertical motion (Option D). This case demonstrates that by anchoring perception through structured knowledge notes, the model can transition from passive pattern matching to an evidence-based verification process, thereby mitigating the identity drift prevalent in standard vision-language reasoning

#### 4.4.2 CASE II: KNOWLEDGE-GUIDED ATTRIBUTE DISCRIMINATION

Figure 4 illustrates a scenario requiring the comparison of plasticity between two rolling balls. In this case, the **Baseline** failure shifts from spatial tracking to causal depth. While the baseline performs a full-motion analysis, it concludes that both balls are identical due to their similar macro-motion. This failure stems from "not knowing what to look for"—the model lacks the specific physical priors to prioritize subtle material-related signals. Conversely, *PhysNote* uses its **Note system** to retrieve task-specific "Tips" from the knowledge base. These notes direct the agent to focus on fine-grained indicators, such as the sharpness of shadows and the degree of spherical retention (identifying a 15% difference in deformation). By externalizing these physical insights, *PhysNote* successfully differentiates the materials. This case shows that our framework allows the model to analyze physically meaningful signals that are typically ignored by baseline.



**Question:** How does the plasticity of the green ball compare to the cyan ball? A. Much less B. Much more  
 C. About the same D. Uncertain  
**Ground Truth:** A

Baseline	Ours
<p><b>Raw Reasoning:</b> To determine the plasticity, let's observe their behavior: 1. <b>Initial Position:</b> Both balls are at the top of their respective ramps. 2. <b>Movement:</b> As the balls roll down, they appear to maintain their spherical shape without deforming significantly. 3. <b>Final Position:</b> Both balls reached the bottom and are resting. Neither ball shows any noticeable deformation or change in shape.</p> <p><b>Conclusion:</b> - Both balls retain spherical shape throughout. - No visible difference in how they deform as they move down and come to rest. Therefore, plasticity is about the same.</p>	<p><b>Raw Reasoning:</b> Tips applied: If the shadow of a ball is sharper and smaller (e.g., green ball), consider lower plasticity, because less material displacement indicates less deformation. If a ball maintains a more spherical shape after interaction, consider lower plasticity.</p> <p><b>Iteration 1:</b> - <i>Hypothesis:</i> Check if green sphere's shadow is sharper/smaller than cyan's in frame 0. Check if green retains a more spherical shape in frame 3 after surface interaction. - <i>Gathered Truths:</i> Green sphere shadow is sharper/smaller. Green sphere retains a more spherical shape (approx. 15% less deformation than cyan).</p> <p><b>Validation:</b> Sharp shadow and spherical retention support lower plasticity for the green sphere.</p>
<b>Extract Answer:</b> C	<b>Extract Answer:</b> A

Figure 4: **Qualitative comparison on physical property reasoning.** The baseline (left) observes no obvious deformation in either ball and concludes that their plasticity is equal (C). Our method (right), guided by retrieved Knowledge Notes on material cues, leverages shadow sharpness and post-interaction shape retention as observable proxies for plasticity. By quantifying the difference in deformation between the two balls, it correctly identifies that the green ball exhibits much less plasticity (A).

## 5 CONCLUSION

We present *PhysNote*, a framework that enhances VLM physical reasoning through externalized Knowledge Notes. By integrating spatio-temporal canonicalization with a hierarchical knowledge structure, we effectively reduce identity drift and reasoning volatility in dynamic scenes. Experimental results on PhysBench show that *PhysNote* achieves 56.68% and 72.86% accuracy on the test and validation sets, respectively.

## REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Anthropic. The Claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, 2024. URL <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL <https://arxiv.org/abs/2502.13923>.
- Vahid Balazadeh, Mohammadmehdi Ataei, Hyunmin Cheong, Amir Hosein Khasahmadi, and Rahul G Krishnan. Physics context builders: A modular framework for physical reasoning in vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7318–7328, 2025.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, et al. How far are we to GPT-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Phys-bench: Benchmarking and enhancing vision-language models for physical world understanding. In *The Thirteenth International Conference on Learning Representations*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Jiafei Duan, Samson Yu, Soujanya Poria, Bihan Wen, and Cheston Tan. Pip: Physical interaction prediction via mental simulation with span selection. In *European Conference on Computer Vision*, pp. 405–421. Springer, 2022.
- Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11474–11484, 2020.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. MANTIS: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024. Best Paper Award.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.
- Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

- Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.
- Machel Reid, Nikolay Savinov, Denis Teplyashin, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. Penetrative ai: Making llms comprehend the physical world. In *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, pp. 1–7, 2024.
- Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024.
- Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaying Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.