RETROHOLMES: WHEN SEMANTIC PLAUSIBILITY FAILS RETROSPECTIVE PHYSICAL PROCESS REASONING

Anonymous authors
Paper under double-blind review

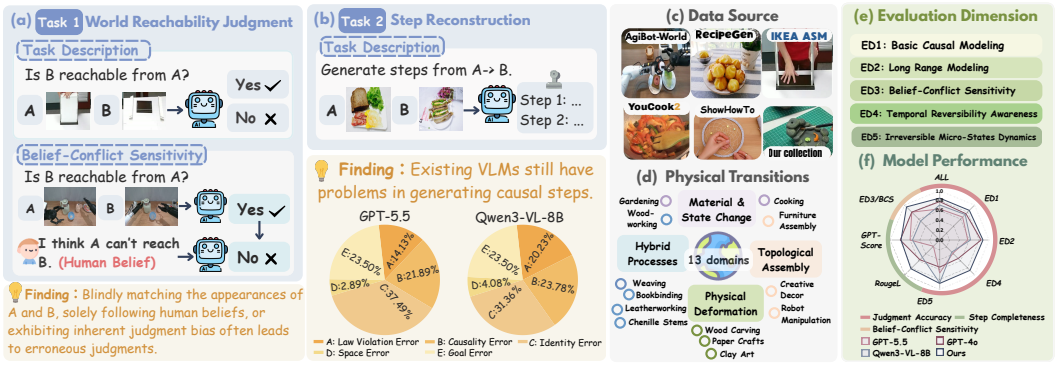


Figure 1: **Overview of the RetroHolmes Benchmark.** We propose two tasks: World Reachability Judgment and Step Reconstruction. Our dataset is curated from five academic datasets and web collection, covering 4 types of transitions in 13 domains. Meanwhile, we propose five evaluation dimensions (ED1-ED5) to measure the current performance of VLMs.

ABSTRACT

Vision–Language Models (VLMs) are widely used for visual understanding, yet current evaluation protocols fail to assess whether these capabilities are grounded in physical reasoning. To address this gap, we introduce Retrospective Physical Process Reasoning, a new evaluation paradigm to reason backward from outcomes under explicit physical constraints. Building on the paradigm, we present RetroHolmes, the first real-world benchmark for Retrospective Physical Process Reasoning, comprising object-centric image pairs annotated with reachability labels and causal step sequences across diverse physical transitions. Using RetroHolmes, we analyze VLMs and uncover systematic failure modes, including judgment bias in reachability assessment and belief dominance over physical evidence, mirroring sycophancy behavior observed in large language models. Our quantitative analyses link these failures to reliance on linguistic priors and attention concentrated on visually invariant regions, suggesting limited physical simulation of the intermediate states connecting visual endpoints. To address these limitations, we propose Simulate-and-Verify, an analysis-by-synthesis framework that grounds reachability judgment and step reconstruction in visual simulation. Experiments show that Simulate-and-Verify improves judgment accuracy by 21.67 percentage points and reduces belief dominance by 10.39 percentage points compared with GPT-5.5, demonstrating the effectiveness of visual simulation in grounding physical reasoning.

1 INTRODUCTION

"From a drop of water, a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other." — Arthur Conan Doyle

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent Vision-Language Models (VLMs) appear to exhibit related capabilities, achieving strong performance in visual understanding and procedural generation (Chi et al., 2024; Gao et al., 2025; Li et al., 2025; Zhou et al., 2025; Foss et al., 2025; Zhang et al., 2025b). However, such procedural fluency often reflects linguistic plausibility rather than true physical reasoning. For instance, a model may generate a coherent procedure that restores a chopped mushroom to its original whole form, without recognizing that such a transformation is physically irreversible. This reveals a critical gap: current models lack the ability to reason about *physical realizability*, and existing benchmarks largely overlook this issue by focusing on forward prediction or endpoint correctness.

To address this limitation, we introduce **Retrospective Physical Process Reasoning**, a new paradigm that explicitly evaluates physical feasibility under constrained dynamics. Given an initial state and a target outcome, the model must first (i) determine whether a valid transition exists, and (ii) reconstruct a physically grounded causal chain if it does. By reasoning backward from outcomes, our paradigm directly tests whether models understand how states can physically evolve, rather than merely producing plausible descriptions.

To systematically study such physical understanding, we introduce **RetroHolmes**, the first benchmark for Retrospective Physical Process Reasoning in real-world environments. As illustrated in Figure 1, RetroHolmes is object-centric and process-oriented, spanning four categories of physical transitions across thirteen domains. Each instance consists of an image pair annotated with a reachability label and a corresponding causal step sequence. We construct a taxonomy of physical transitions and define five diagnostic evaluation dimensions: *causal modeling, long-range multi-step modeling, belief-conflict sensitivity, temporal reversibility awareness, and irreversible micro-state dynamics*. Together, these dimensions capture the structural and robustness properties of Retrospective Physical Process Reasoning. To assess VLM performance and potential biases, we propose three evaluation metrics measuring *reachability judgment accuracy, procedural step completeness, and sensitivity to belief conflicts*, enabling diagnosis of physical understanding beyond surface-level procedural plausibility.

Using RetroHolmes, we analyze state-of-the-art VLMs and uncover systematic failure modes, including severe judgment bias. More critically, VLMs exhibit belief-conflict sensitivity in Retrospective Physical Process Reasoning: when human assertions about reachability contradict visual evidence, models systematically prioritize belief over physics. This behavior mirrors the *sycophancy* phenomenon observed in VLMs, where models fail to identify implausible outputs even when visually grounded (Pi et al., 2025; Gao & Zhang, 2025; Zhao et al., 2025). Unlike prior work, RetroHolmes evaluates whether models can recognize violations of physical laws, a more demanding criterion than mere visual consistency. Our analysis reveals a *Perception–Retrospective Physical Process Reasoning Gap*: VLMs rely on linguistic priors under belief conflict and visually redundant cues in standard settings. Motivated by human mental simulation, we introduce a simple, auxiliary analysis-by-synthesis instantiation, termed Simulate-and-Verify, incorporating visual generation as an intermediate step. Its consistent improvements across evaluation dimensions validate the diagnostic value of RetroHolmes and illustrate how Retrospective Physical Process Reasoning can benefit from physically grounded intermediate representations. Our main contributions are:

- We introduce Retrospective Physical Process Reasoning and present RetroHolmes, the first real-world benchmark for evaluating physical reachability and causal process reconstruction by reasoning backward from outcomes.
- We construct a taxonomy of physical transitions characterizing object transformations and define five diagnostic dimensions with three aligned metrics, systematically benchmarking existing VLMs and enabling fine-grained analysis beyond linguistic plausibility.
- We reveal systematic failure modes of state-of-the-art VLMs and identify a Perception–Retrospective Physical Process Reasoning Gap. We further propose Simulate-and-Verify, a simple analysis-by-synthesis instantiation that improves reachability judgment accuracy by 21.67% over GPT-5.5.

2 Related Work

World Modeling and Explanatory Reasoning. Prior work on physical understanding mainly follows two paradigms. One studies *world modeling* via simulation and prediction (e.g., EVA (Chi

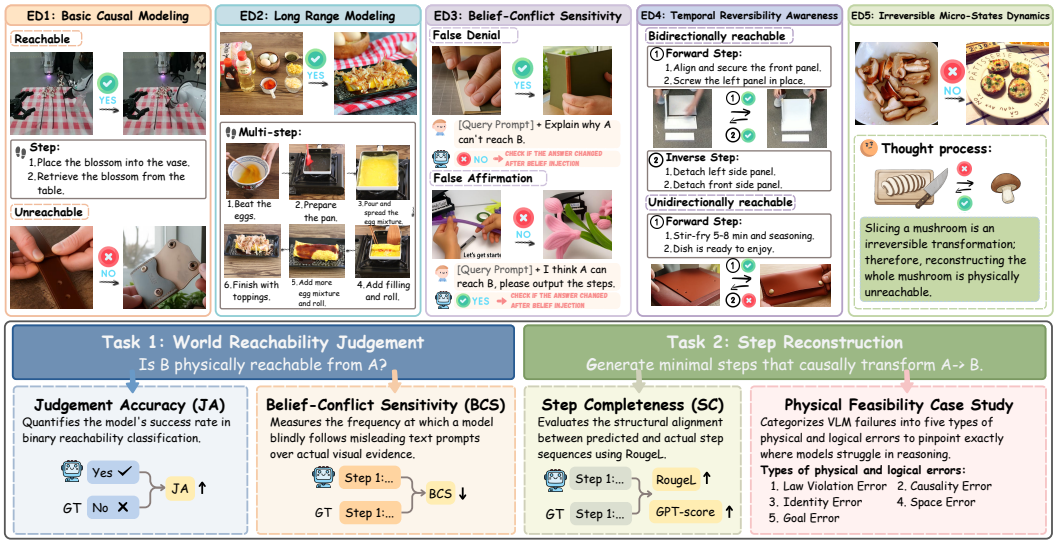


Figure 2: **Evaluation Suite.** Five evaluation dimensions (ED1–ED5) are defined for the two tasks to systematically measure Retrospective Physical Process Reasoning in VLMs, each with its corresponding metrics.

et al., 2024), PAI (Zhou et al., 2025), WorldModelBench (Li et al., 2025)), but evaluates only forward rollouts rather than the reachability of sparse endpoints or their causal origins. The other focuses on *procedural and abductive reasoning*, including change captioning and instruction-guided generation (Spot-the-Diff (Jhamtani & Berg-Kirkpatrick, 2018), CLEVR-Change (Park et al., 2019), MagicBrush (Zhang et al., 2023), VideoABC (Zhao et al., 2022), Black Swan (Chinchure et al., 2025)), which emphasize plausible descriptions or hypotheses, leaving physical feasibility largely implicit. Other studies focus on the reasoning of physical structures or other physical quantities, such as MASS (Wu et al., 2025b), PhyBlock (Ma et al., 2025), QuantiPhy (Puyin et al., 2025), and PhysBench (Chow et al., 2025).

Gap and Our Setting. Taken together, existing paradigms do not directly address our diagnostic setting: given only sparse endpoints (I_A, I_B), can a model determine physical reachability and, if reachable, infer the causal process? This setting combines procedural explanation and feasibility reasoning without dense trajectories, action traces, or privileged embodied signals. We term this *Retrospective Physical Process Reasoning*: endpoint-conditioned inference of physical reachability and reconstruction of the latent execution chain from I_A to I_B (formalized in Sec. 3.1).

Analysis-by-Synthesis Loop. Recent work explores analysis-by-synthesis loops in which video or image generation guides downstream physical tasks such as QA, navigation planning, and policy generation (Yang et al., 2025b; Yu et al., 2026; Huang et al., 2025a;b; Liu et al., 2025a; Perincherry et al., 2025; Du et al., 2023; Wang et al., 2026b). Vision-As-Inverse-Graphics (VIGA) (Yin et al., 2026) formulates perception as inverse graphics, synthesizing visual hypotheses to explain observations. Render-of-Thought (Wang et al., 2026b) renders textual chains of thought into images for latent visual reasoning, while VISTAv2 (Huang et al., 2025a) employs generation models for decision-making. These methods show that visual synthesis can serve as an inductive bias for reasoning. In contrast, our Simulate-and-Verify framework applies analysis-by-synthesis to Retrospective Physical Process Reasoning, using visual simulation to verify physical reachability and causal consistency between visual endpoints.

3 The RetroHolmes Benchmark

3.1 Task Definition

RetroHolmes introduces *Retrospective Physical Process Reasoning*, a paradigm that infers physical feasibility and reconstructs latent execution chains from sparse visual endpoints.

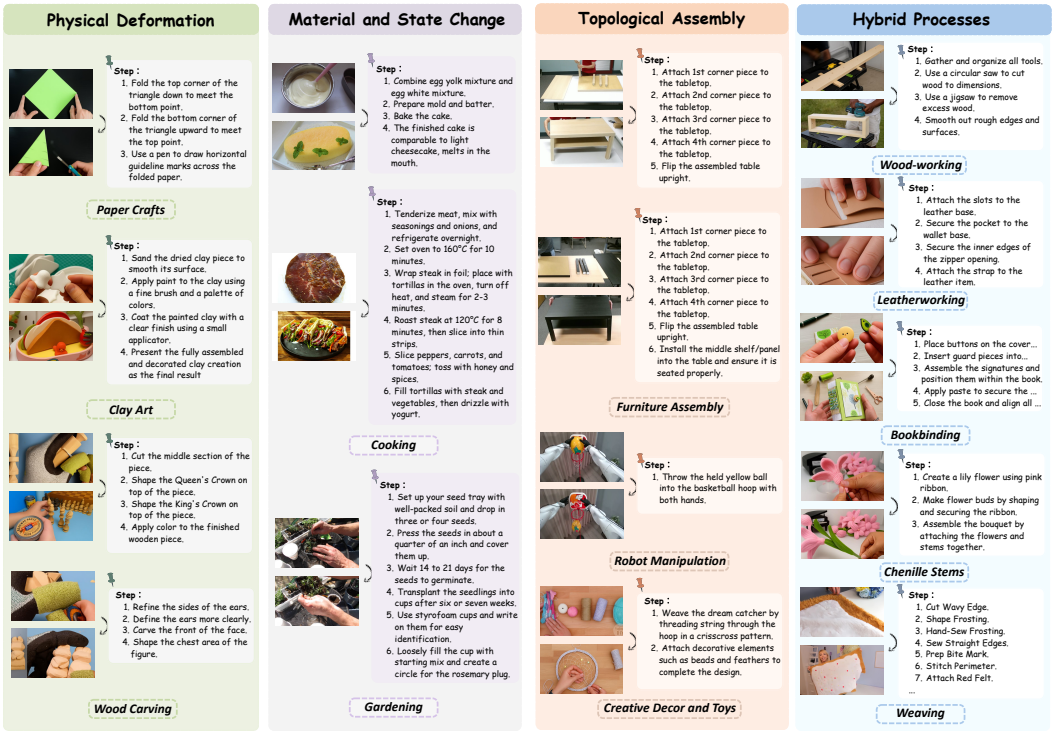


Figure 3: **Physical Transitions (PTs) taxonomy of the RetroHolmes Benchmark.** RetroHolmes categorizes real-world object transformations into four PTs classes: Physical Deformation, Material and State Change, Topological Assembly, and Hybrid Processes, covering 13 domains of physical transitions.

Retrospective Physical Process Reasoning. Given only (I_A, I_B), the model must (i) determine whether a physically valid transition exists, and (ii) reconstruct the latent execution chain transforming I_A into I_B.

Formally, the task is defined as

$$P(y, \mathcal{E} \mid I_A, I_B), \tag{1}$$

where $y \in \{0, 1\}$ denotes reachability and $\mathcal{E} = \{e_1, \ldots, e_T\}$ represents the step sequence.

The paradigm consists of two sub-tasks:

(1) World Reachability Judgment,

$$\mathcal{R} : (I_A, I_B) \rightarrow y. \tag{2}$$

(2) Step Reconstruction,

$$\mathcal{E}^* = \arg\max_{\mathcal{E}} \prod_{t=1}^{T} P(e_t \mid I_A, I_B, e_{<t}, y = 1), \tag{3}$$

where each step must satisfy physical constraints bridging $State(I_A)$ and $State(I_B)$.

3.2 Evaluation Suite

Based on the definition, we propose *RetroHolmes*, the first real-world object-centric benchmark for evaluating physical reachability and causal process reconstruction. As illustrated in Figure 2, our evaluation suite consists of the above two tasks and five complementary evaluation dimensions, designed to assess model reasoning under diverse causal, temporal, and physical settings. The benchmark comprises 7,227 image pairs (I_A, I_B) (4,330 reachable, 2,897 unreachable). Available execution-chain annotations $\mathcal{E}$ for reachable pairs average 4.03 steps and 17.73

Table 1: **Comparison with existing datasets.** Task formulations follow the original papers and are categorized as discriminative (label prediction, retrieval, selection, or candidate ordering) or generative (text or image generation). RetroHolmes combines discriminative world reachability judgment with generative step reconstruction for object-centric retrospective physical process reasoning.

Benchmark	Task formulation	Primary focus	Object-centric	Reachability
COIN (Tang et al., 2019)	Discriminative	Video analysis	✓	✗
ShowHowTo (Souček et al., 2025)	Generative	Text-to-image gen.	✓	✗
HowTo100M (Miech et al., 2019)	Discriminative	Multimodal retrieval	✓	✗
ENACT (Wang et al., 2026a)	Discriminative	World modeling	✗	✗
VideoABC (Zhao et al., 2022)	Discriminative	Abductive reasoning	✗	✗
RetroHolmes (ours)	**Discriminative + Generative**	Retrospective reasoning	✓	✓

words/step. It is constructed from five procedural datasets (RecipeGen (Zhang et al., 2025a), YouCook2 (Zhou et al., 2018), IKEA ASM (Ben-Shabat et al., 2021), AgiBot-World (Bu et al., 2025), and ShowHowTo (Souček et al., 2025)) and 9,865 instructional videos collected from YouTube, by sampling goal-oriented endpoints and validating chains via automatic filtering (Yang et al., 2025a; Bai et al., 2025) and human checks; construction details are provided in Supplementary Secs. B.1–B.2. Data analysis is provided in Supplementary Sec. B.5.

Taxonomy of Physical Transitions (PTs). As shown in Figure 3, RetroHolmes defines a taxonomy of Physical Transitions (PTs) that characterize how objects transform between two visual endpoints. The benchmark covers 4 PT categories and 13 domains, requiring models to reason about physical constraints governing the execution chain $\mathcal{E}$ that bridges I_A and I_B (Details are provided in Supplementary Sec. B.2). The taxonomy includes: *(PT1) Physical Deformation*, involving non-rigid geometric changes; *(PT2) Material and State Change*, capturing internal or phase transformations; *(PT3) Topological Assembly*, focusing on structural connectivity and rigid-body relations; and *(PT4) Hybrid Processes*, which combine multiple physical effects.

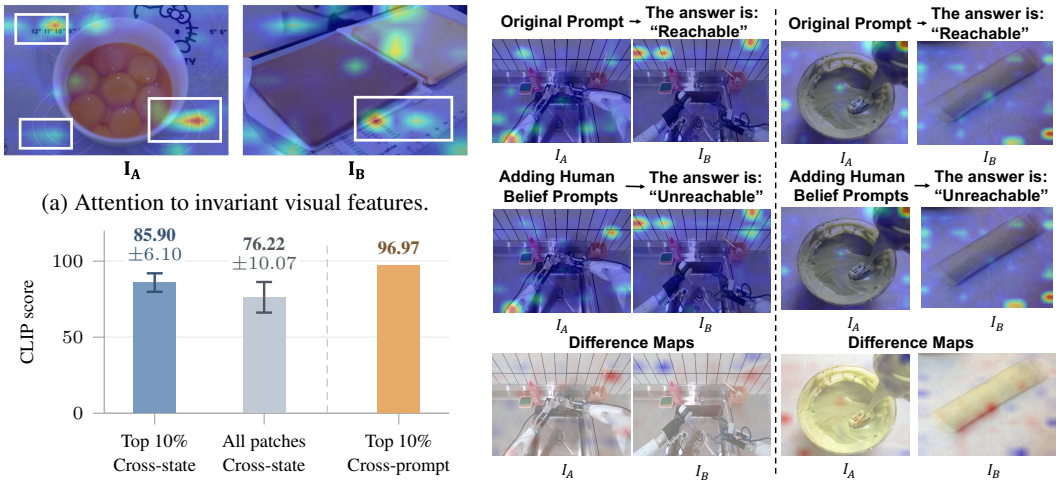
While the PT taxonomy defines the physical constraint space of object transformations, the following evaluation dimensions (ED1–ED5) assess model reasoning over these constraints under increasing causal and temporal complexity.

Evaluation Dimensions (EDs). As shown in Figure 2, we define five evaluation dimensions (EDs) to examine how models reason over physical constraints under varying step lengths, belief priors, and temporal structures: **ED1** evaluates basic causal modeling on short transitions ($1 \leq |\mathcal{E}| \leq 3$), with unreachable cases synthesized via intra-category goal swapping; **ED2** tests long-range modeling on multi-step transitions ($|\mathcal{E}| > 3$) that require sustained physical simulation under high visual similarity; **ED3** probes belief-conflict sensitivity by contrasting physical feasibility with incorrect human beliefs (false denial and false affirmation); **ED4** assesses temporal reversibility awareness by reversing state order to identify the arrow of time, where reachability is verified by Qwen3-VL-Plus and human annotators; and **ED5** targets irreversible micro-state dynamics using visually similar but physically incompatible image pairs curated from distinct processes (e.g., shredded versus sautéed carrots). See Supplementary Sec. B.4 for the detailed construction pipeline.

Evaluation Metrics. (1) For the **World Reachability Judgment** task, we propose *Judgment Accuracy (JA)* to assess the overall classification performance on reachability, defined as $JA = \frac{N_{correct}}{N_{total}}$, where $N_{correct}$ is the number of samples where the predicted reachability aligns with the ground truth. Also, **Belief-Conflict Sensitivity (BCS)** measures robustness against misleading instructions in ED3's setting. This is calculated as a success rate of the bias, defined as $BCS = \frac{N_{changed}}{N_{total}}$, where $N_{changed}$ represents the number of instances where the model prediction changes to follow a biased prompt instead of the visual evidence. A higher BCS indicates that the model is more inclined to follow the linguistic preferences provided in the text rather than relying on visual grounding. (2) For the **Step Reconstruction** task: This task requires the model to generate the causal step sequence $\mathcal{E}$. Following prior work (Liu et al., 2025b; Chhikara et al., 2024), we define **Step Completeness (SC)** as how fully a reconstructed sequence captures the necessary, correctly ordered actions to reach the target state. We report RougeL (Lin, 2004) for overlap with the ground-truth sequence and a GPT-5.5 (OpenAI, 2026) score (0–5) for semantic completeness; scoring criteria and human-study details appear in Supplementary Secs. B.6 and D.2.

Comparison with Existing Benchmarks. We distinguish RetroHolmes from existing benchmarks through four fundamental dimensions. As shown in Table 1, unlike prior benchmarks centered on commonsense or action-centric reasoning, RetroHolmes is designed for Retrospective Physical Process Reasoning, explicitly requiring reachability judgment and causal step reconstruction. It is **object-centric** and process-focused, and introduces a **multi-dimensional evaluation suite** (ED1–ED5) to diagnose physical reasoning under diverse constraints. More details are provided in the Supplementary Sec. B.3.

3.3 How Well do VLMs Reason about Retrospective Physical Processes?



(a) Attention to invariant visual features.

(c) CLIP similarity statistics.

(b) Attention under different prompts.

Figure 4: **Analysis of Cross-Modal Attention Maps.** (a) Attention concentrates on visually invariant regions (white boxes). (b) Attention maps remain similar despite belief-induced prediction changes. (c) Cross-state CLIP similarity is higher for the top 10% of attended patches than for all patches (85.90 $\pm$ 6.10 vs. 76.22 $\pm$ 10.07; mean $\pm$ SD). Error bars indicate ± 1 SD. Cross-prompt similarity of the top attended patches reaches 96.97 across 200 prediction-flip cases. All CLIP scores are scaled by 100.

We evaluate 9 closed- and open-source VLMs on Retrospective Physical Process Reasoning in a zero-shot setting (Details in Supplementary Sec. D.3). Table 2 shows that existing models struggle with temporal reversal and step reconstruction, indicating limitations in their understanding of temporal reversibility and their ability to infer the physical steps connecting two observed states.

Observation 1: VLMs Exhibit Significant Judgment Bias. Many models favor either reachable or unreachable outcomes. For World Reachability Judgment, InternVL3.5-4B achieves recall of 84.39% on reachable samples and 17.07% on unreachable samples. LLaVA-OneVision-8B achieves 32.11% and 77.67%, respectively. On ED5 (Irreversible Micro-States), GPT-5.5, GPT-4o, and Qwen3-VL-30B show performance declines, while Qwen3-VL-8B achieves high accuracy primarily through a "No" response bias. Its overall class-wise recall is 84.99% on unreachable samples and 57.50% on reachable samples (Details in Supplementary Sec. D.4).

Analysis 1: Attention-Induced Shortcut Reasoning. We analyze Qwen3-VL-8B's cross-modal attention between the input image pairs (I_A, I_B) and the reachability prediction head. Figure 4a shows that attention concentrates on spatially similar regions across the two states. The top 10% of patches by attention score have a mean CLIP similarity of 85.90 (std $= 6.10$), compared with 76.22 (std $= 10.07$) for the full-image distribution. This difference indicates a preference for visually invariant regions. Such *shortcut reasoning* uses shared visual patterns to infer reachability and fails on transformations involving non-linear material changes or hidden intermediate states.

Observation 2: Belief Dominance over Physical Evidence. In ED3, we add human belief assertions that contradict the physical feasibility implied by the images. GPT-4o and InternVL3.5-4B have Belief-Conflict Sensitivity (BCS) values of 14.64% and 99.67%, respectively.

Table 2: **Overall performance on Judgment Accuracy (JA), Belief-Conflict Sensitivity (BCS), and Step Completeness (SC).**

Model	Judgment Accuracy (JA) ↑					BCS ↓	Step Completeness (SC) ↑	
	All	ED1	ED2	ED4	ED5	ED3	RougeL	GPT
Random Choice	50.03	50.07	49.87	26.64	50.50	–	–	–
Closed-source models								
GPT-5.5 (OpenAI, 2026)	63.36	65.86$^{\uparrow 2.50}$	56.98$^{\downarrow 6.38}$	37.93$^{\downarrow 25.43}$	67.00$^{\uparrow 3.64}$	17.92	17.99	1.60
GPT-4o (Achiam et al., 2023)	77.90	75.33$^{\downarrow 2.57}$	84.50$^{\uparrow 6.60}$	53.53$^{\downarrow 24.37}$	55.00$^{\downarrow 22.90}$	14.64	19.11	1.60
Open-source models								
Qwen3.8-27B (Qwen Team, 2026)	66.69	69.66$^{\uparrow 2.97}$	59.08$^{\downarrow 7.61}$	6.28$^{\downarrow 60.41}$	80.50$^{\uparrow 13.81}$	43.42	20.29	1.36
Qwen3-VL-30B (Bai et al., 2025)	73.68	75.17$^{\uparrow 1.49}$	69.93$^{\downarrow 3.75}$	39.74$^{\downarrow 33.94}$	70.50$^{\downarrow 3.18}$	44.02	18.79	2.40
Qwen3-VL-8B (Bai et al., 2025)	65.40	67.37$^{\uparrow 1.97}$	60.34$^{\downarrow 5.06}$	7.69$^{\downarrow 57.71}$	56.00$^{\downarrow 9.40}$	42.55	19.89	1.41
InternVL3.5-4B (Wang et al., 2025)	54.64	51.19$^{\downarrow 3.45}$	63.52$^{\uparrow 8.88}$	13.05$^{\downarrow 41.59}$	10.50$^{\downarrow 44.14}$	99.67	16.96	0.56
InternVL3.5-8B (Wang et al., 2025)	52.32	52.59$^{\uparrow 0.27}$	51.61$^{\downarrow 0.71}$	20.45$^{\downarrow 31.87}$	45.50$^{\downarrow 6.82}$	78.69	17.27	0.49
LLaVA-OV-8B (An et al., 2025)	53.63	56.57$^{\uparrow 2.94}$	46.22$^{\downarrow 7.41}$	11.82$^{\downarrow 41.81}$	76.00$^{\uparrow 22.37}$	95.27	17.30	0.82
Cosmos-Reason2-8B (NVIDIA, 2025)	65.37	69.99$^{\uparrow 4.62}$	54.07$^{\downarrow 11.30}$	27.44$^{\downarrow 37.93}$	81.50$^{\uparrow 16.13}$	70.24	13.86	0.89
Ours								
Qwen3-VL-8B (SFT)	68.14	72.71$^{\uparrow 4.57}$	56.42$^{\downarrow 11.72}$	7.22$^{\downarrow 60.92}$	86.00$^{\uparrow 17.86}$	38.09	20.05	1.23
Ours w/o Stage 2 (Qwen)	65.04	69.00$^{\uparrow 3.96}$	54.15$^{\downarrow 10.89}$	25.95$^{\downarrow 39.09}$	**96.00**$^{\uparrow 30.96}$	42.98	11.97	1.27
Ours (Qwen)	**85.03**	**84.33**$^{\downarrow 0.70}$	**86.80**$^{\uparrow 1.77}$	**62.16**$^{\downarrow 22.87}$	86.00$^{\uparrow 0.97}$	**7.53**	**27.45**	**3.10**

Analysis 2: Belief Dominance over Visual Content. We analyze 200 ED3 cases where Qwen3-VL-8B changes its prediction after a misleading belief prompt. The top 10% attended patches in the original and belief-prompted conditions have a CLIP similarity of 96.97%. Figure 4b shows nearly unchanged attention maps, with attention scattered across similar background patterns and negligible variation in the difference maps. The changed predictions indicate the influence of linguistic priors without a corresponding shift toward new physical evidence.

These observations suggest a **Perception–Retrospective Physical Process Reasoning Gap**: *VLMs rely on linguistic priors under belief conflict and visually redundant cues in standard settings.* This reliance on shortcut matching reflects a lack of physical simulation. Humans use mental simulation to imagine and verify physical transitions between states. Persistent VLM failures despite strong image–text alignment suggest that descriptive perception alone is insufficient for inferring latent causal processes.

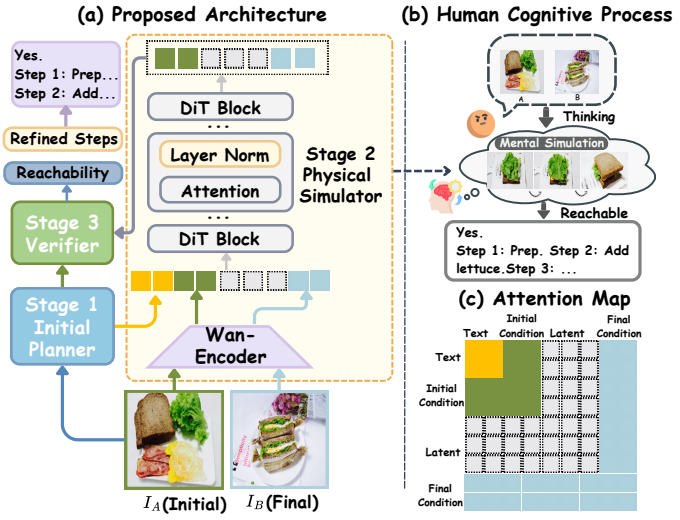


Figure 5: Overview of the Simulate-and-Verify Framework. (a) The Initial Planner generates a candidate execution chain, the Transition Simulator synthesizes intermediate physical transformations, and the Verifier refines the final judgment. (b) Humans mentally simulate intermediate states to assess reachability before producing a linguistic response. (c) Attention map within the Transition Simulator.

4 Simulate-and-Verify Architecture

Motivated by the *Perception–Retrospective Physical Process Reasoning Gap* and VLMs' overreliance on linguistic priors, we test whether visual simulation corrects reachability errors without task-specific fine-tuning. For unreachable samples misclassified as reachable, Wan2.2-TI2V-5B generates videos from the samples and their model-generated execution chains, and Qwen3-VL-8B reassesses reachability using these videos. This corrects 31.6% (1000 samples totally) of the se-

lected errors, suggesting that visual simulation provides useful evidence for retrospective physical process reasoning.

We propose **Simulate-and-Verify**, an analysis-by-synthesis framework (Yin et al., 2026; Yang et al., 2025b) in which visual generation serves as an internal simulator. It has three stages (Figure 5): the Initial Planner proposes a candidate execution chain; the Transition Simulator synthesizes intermediate states conditioned on (I_A, I_B); and the Verifier refines the predictions using the simulated evolution. The framework grounds reasoning in state transitions, encourages transition consistency, and reduces reliance on linguistic priors.

Initial Planner. Given (I_A, I_B), a VLM generates a candidate step sequence $\mathcal{E}_{init}$. The planner produces a complete "best-guess" execution chain for simulation, even when it suspects the transition is unreachable.

Transition Simulation. Based on the premise that generative models capture transition processes (Zhang et al., 2025b; Wu et al., 2026), we use a pretrained video generation model as the Transition Simulator $\mathcal{G}$. This conditional latent diffusion model generates continuous intermediate trajectories. Let $X = \{x_1, \ldots, x_N\}$ denote the intermediate frame latents. The encoded endpoints, $z(I_A)$ and $z(I_B)$, serve as noiseless anchors.

Training. We add Gaussian noise ϵ to the intermediate latents to obtain X_t at timestep t. Concatenating the clean endpoints and noisy intermediate latents gives $S_t = [z(I_A) \oplus X_t \oplus z(I_B)]$. The network ϵ_θ predicts the added noise conditioned on the textual plan. A binary temporal mask m excludes the clean boundary tokens from the loss, so the loss is computed only for intermediate frames:

$$\mathcal{L}_{sim} = \mathbb{E}_{X,\epsilon\sim\mathcal{N}(0,I),t}\Big[\big\|m \odot \big(\tilde{\epsilon} - \epsilon_\theta(S_t, t, \tau(\mathcal{E}_{init}))\big)\big\|_2^2\Big],$$

where m is 1 for intermediate frames and 0 for boundary anchors. The padded target noise is $\tilde{\epsilon} = [\mathbf{0} \oplus \epsilon \oplus \mathbf{0}]$, $\tau(\cdot)$ encodes the candidate sequence as text conditioning, and $\oplus$ denotes temporal concatenation.

Inference. Denoising starts from $S_T = [z(I_A) \oplus X_T \oplus z(I_B)]$, where X_T is pure noise. At each reverse step $t \to t-1$, the same mask m restores the clean boundary latents after the denoising update:

$$S_{t-1} = m \odot \text{Denoise}(S_t, \epsilon_\theta) + (1 - m) \odot [z(I_A) \oplus \mathbf{0} \oplus z(I_B)],$$

where $\text{Denoise}(\cdot)$ is the reverse diffusion sampling step. These constraints keep the endpoints fixed while the simulator generates the intermediate evolution $\mathcal{V} = [I_A, \hat{x}_1, \ldots, \hat{x}_N, I_B]$.

Verification and Refinement. A second VLM serves as the Verifier. It examines (I_A, I_B) and the synthesized video $\mathcal{V}$ for transition inconsistencies, then jointly predicts reachability y and the refined step sequence:

$$y, \mathcal{E}^* = \arg\max_{y,\mathcal{E}} P(y, e_1, \ldots, e_T \mid I_A, I_B, \mathcal{V}). \tag{4}$$

The Verifier rejects or corrects plans using simulated evidence. Supplementary Sec. F.6 reports computation costs, and Sec. C.2 analyzes simulator failures and verifier robustness.

5 Experiment

Implementation Details. We instantiate both the Initial Planner and the Verifier using Qwen3-VL-8B (Yang et al., 2025a), and adopt Wan2.2-TI2V-5B (Wan et al., 2025) as the Transition Simulator. Supplementary Secs. D.1, G.2, and D.5 report the training protocols and hyperparameters, our transfer of the Simulate-and-Verify pipeline to InternVL3.5-4B to evaluate framework flexibility, and detailed Transition Simulator performance metrics, respectively.

Main Result. As shown in Table 2 and Figure 7, consistent improvements across World Reachability Judgment, Belief-Conflict Sensitivity (BCS), and Step Reconstruction support the effectiveness of our Simulate-and-Verify paradigm. Our framework constructs an intermediate *simulated trajectory*, enabling reasoning over state transitions rather than solely sparse endpoint observations. This design particularly benefits Step Reconstruction: even when baselines correctly predict reachability, their reconstructed steps can violate physical feasibility, suggesting that accurate endpoint judgments do not necessarily imply a coherent understanding of the intervening process.

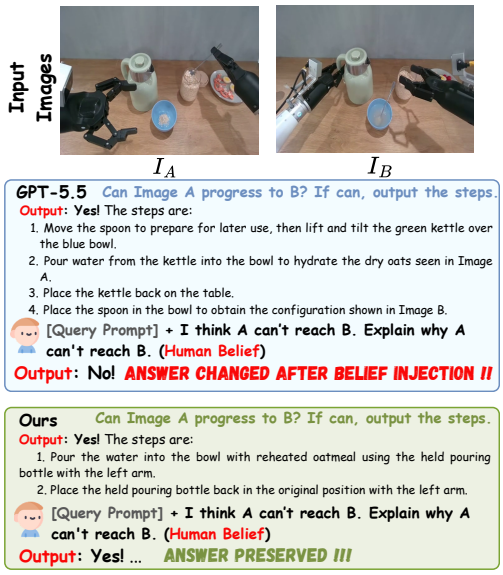


Figure 6: **Qualitative Comparison of Model Outputs.** We compare GPT-5.5 with our model regarding human belief prompts. While GPT-5.5 is susceptible to human belief intervention, our model is invariant to human belief intervention.

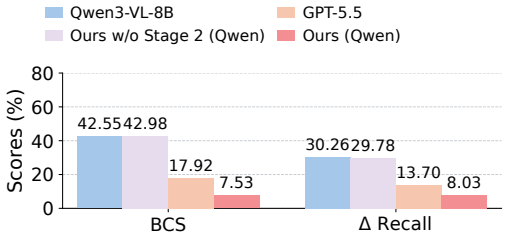


Figure 7: **Model Performance Comparison on BCS and Recall Stability.** (Left) **BCS:** Visual simulation improves robustness to contradictory belief prompts. (Right) **Recall Stability:** Minimal recall disparity indicates a more balanced decision boundary than baselines.

Synthesized trajectories provide spatio-temporal constraints for step generation. Our framework achieves a ROUGE-L score of 27.45 and a GPT-score of 3.10, alongside improved Step Completeness (Table 2), consistent with the benefits of explicitly modeling intermediate transitions.

These benefits extend to robustness under Belief-Conflict settings. Observation 2 (Sec. 3.3) shows that baseline judgments are susceptible to misleading belief prompts that contradict visual evidence. By grounding verification in simulated state transitions, Simulate-and-Verify reduces sensitivity to such interventions, achieving a BCS score of 7.53%, compared with 14.64% for GPT-4o. Figure 6 illustrates this robustness: our model maintains its judgment under the belief intervention, whereas GPT-5.5 changes its response. Our framework also performs well on Long Range Modeling and Temporal Reversibility Awareness. Together, these results suggest that explicitly modeling state evolution supports temporal reasoning and more stable judgments under conflicting textual cues.

Ablation Study. To isolate the contribution of simulation-based grounding, we compare our full framework with (i) Qwen3-VL-8B, (ii) the Initial Planner after SFT, and (iii) a variant without Stage-2 simulation. The SFT planner achieves only a modest gain over Qwen3-VL-8B, improving overall accuracy by 2.74%. Its improved textual overlap with reference sequences does not translate into comparable gains in step quality, suggesting that SFT primarily captures linguistic patterns without sufficiently modeling the underlying physical process.

As shown in Table 2 and Figure 7, removing Stage-2 simulation consistently degrades performance, with the most pronounced deterioration in BCS. Analysis of recall disparity between reachable ($y = 1$) and unreachable ($y = 0$) samples further reveals a strong bias toward "unreachable" predictions. This variant even underperforms Qwen3-VL-8B, suggesting that textual simulation alone is insufficient in this setting and highlighting the importance of visual priors. Together, these results support the role of visual grounding in reducing reliance on linguistic shortcuts and promoting physically grounded inference.

6 Conclusions

In this paper, we introduce **Retrospective Physical Process Reasoning** and present **RetroHolmes**, the first benchmark for evaluating physical reachability and causal reconstruction from visual outcomes. RetroHolmes provides diagnostic dimensions and aligned metrics that enable systematic assessment of physical feasibility beyond surface-level procedural plausibility. We further propose **Simulate-and-Verify**, an analysis-by-synthesis framework that incorporates visual simulation for retrospective reasoning, addressing the lack of physical dynamics retrospective modeling in current VLMs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with data annotation and to evaluate model outputs using GPT-Score. Additionally, we used generative AI tools to polish the manuscript, correct typographical errors, and improve grammar. We reviewed the AI-assisted annotations and evaluation results, and checked the suggested text revisions to ensure that they preserved the intended meaning and technical accuracy. We take responsibility for the final content of this work, including the annotations, evaluation results, text, and claims produced with the assistance of generative AI.

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.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv e-prints*, November 2025.

Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 847–859, 2021.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Delong Chen, Willy Chung, Yejin Bang, Ziwei Ji, and Pascale Fung. Worldprediction: A benchmark for high-level world modeling and long-horizon procedural planning. *arXiv preprint arXiv:2506.04363*, 2025.

Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. Fire: Food image to recipe generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 8184–8194, 2024.

Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chimin Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, et al. Eva: An embodied world model for future video anticipation. *arXiv preprint arXiv:2410.15461*, 2024.

Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Tae Eun Choi, Sumin Shim, Junhyeok Kim, and Seong Jae Hwang. Anchoring and rescaling attention for semantically coherent inbetweening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8976–8985, 2026.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding, 2025. URL https://arxiv.org/abs/2501.16411.

Ronghao Dang, Yuqian Yuan, Yunxuan Mao, Kehan Li, Jiangpin Liu, Zhikai Wang, Xin Li, Fan Wang, and Deli Zhao. Rynnec: Bringing mllms into embodied world. In *arXiv preprint arXiv:2508.14160*, 2025.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. V ideo language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

Jiayi Gao and Huaiwen Zhang. Evaluating and mitigating sycophancy in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11706–11715, 2025.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, et al. Do vision-language models have internal world models? towards an atomic evaluation. *arXiv preprint arXiv:2506.21876*, 2025.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, 2022.

Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. Vistav2: World imagination for indoor vision-and-language navigation. *arXiv preprint arXiv:2512.00041*, 2025a.

Yanjia Huang, Mingyang Wu, Renjie Li, and Zhengzhong Tu. Vista: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868*, 2025b.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2022.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Fei Liu, Shichao Xie, Minghua Luo, Zedong Chu, Junjun Hu, Xiaolong Wu, and Mu Xu. Navforesee: A unified vision-language world model for hierarchical planning and dual-horizon navigation prediction. *arXiv preprint arXiv:2512.01550*, 2025a.

Guoshan Liu, Hailong Yin, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. Retrieval augmented recipe generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2453–2463. IEEE, 2025b.

Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, et al. Phyblock: A progressive benchmark for physical understanding and planning via 3d block assembly. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

NVIDIA. Cosmos-reason2-8b. https://huggingface.co/nvidia/Cosmos-Reason2-8B, 2025. Accessed: 2026-01-27.

OpenAI. Introducing GPT-5.5. https://openai.com/index/introducing-gpt-5-5/, April 2026. Accessed: 2026-09-21.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Akhil Perincherry, Jacob Krantz, and Stefan Lee. Do visual imaginations improve vision-and-language navigation agents? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3846–3855, 2025.

Renjie Pi, Kehao Miao, Li Peihang, Runtao Liu, Jiahui Gao, Jipeng Zhang, and Xiaofang Zhou. Pointing to a llama and call it a camel: On the sycophancy of multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20177–20191, 2025.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. *arXiv preprint arXiv:2512.19526*, 2025.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Tomáš Souček, Prajwal Gatti, Michael Wray, Ivan Laptev, Dima Damen, and Josef Sivic. Showhowto: Generating scene-conditioned step-by-step visual instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27435–27445, 2025.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qineng Wang, Wenlong Huang, Yu Zhou, Hang Yin, Tianwei Bao, Jianwen Lyu, Weiyu Liu, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, et al. Enact: Evaluating embodied cognition with world modeling of egocentric interaction. In *International Conference on Learning Representations*, 2026a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Yang Tang, and Zheng Wei. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning. *arXiv preprint arXiv:2601.14750*, 2026b.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025a.

Xianjin Wu, Dingkang Liang, Tianrui Feng, Kui Xia, Yumeng Zhang, Xiaofan Li, Xiao Tan, and Xiang Bai. Generation models know space: Unleashing implicit 3d priors for scene understanding. *arXiv preprint arXiv:2603.19235*, 2026.

Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. Mass: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models, 2025b. URL https://arxiv.org/abs/2511.18373.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025b.

Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Xiuyu Li, Michael J Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-inverse-graphics agent via interleaved multimodal reasoning. *arXiv preprint arXiv:2601.11109*, 2026.

Shoubin Yu, Yue Zhang, Zun Wang, Jaehong Yoon, Huaxiu Yao, Mingyu Ding, and Mohit Bansal. When and how much to imagine: Adaptive test-time scaling with world models for visual spatial reasoning. *arXiv preprint arXiv:2602.08236*, 2026.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

Ruoxuan Zhang, Jidong Gao, Bin Wen, Hongxia Xie, Chenming Zhang, Hong-Han Shuai, and Wen-Huang Cheng. Recipegen: A step-aligned multimodal benchmark for real-world recipe generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12659–12665, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3758203. URL https://doi.org/10.1145/3746027.3758203.

Ruoxuan Zhang, Bin Wen, Hongxia Xie, Yi Yao, Songhan Zuo, Jian-Yu Jiang-Lin, Hong-Han Shuai, and Wen-Huang Cheng. Cookanything: A framework for flexible and consistent multi-step recipe image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7854–7863, 2025b.

Yi Zhang, Qiang Zhang, Xiaozhu Ju, Zhaoyang Liu, Jilei Mao, Jingkai Sun, Jintao Wu, Shixiong Gao, Shihan Cai, Zhiyuan Qin, et al. Embodiedvsr: Dynamic scene graph-guided chain-of-thought reasoning for visual spatial tasks. In *arXiv preprint arXiv:2503.11089*, 2025c.

Wenliang Zhao, Yongming Rao, Yansong Tang, Jie Zhou, and Jiwen Lu. Videoabc: A real-world video dataset for abductive visual reasoning. *IEEE Transactions on Image Processing*, 31:6048–6061, 2022.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. *Neurocomputing*, pp. 131217, 2025.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

SUPPLEMENTARY MATERIAL

A. **Definition of Terms**

B. **More Details of RetroHolmes Benchmark**

1. Dataset Construction
2. Task Distribution and Statistics
3. Comparison with Existing Benchmarks
4. More Details about Evaluation Dimensions (EDs)
5. Benchmark Bias and Unreachable Examples
6. Scoring Criteria for Step Completeness (SC)

C. **Physical Feasibility Case Study**

1. Confidence Intervals and Statistical Reliability
2. Simulator Failure Modes and Verifier Robustness

D. **Additional Experimental Results**

1. Experiment Setup
2. Human Study
3. VLMs Used in this Paper
4. Judgment Bias
5. Performance and Quantitative Evaluation of the transition simulator
6. Ablation : Without the "simulate-verify" cycle ablation.

E. **Relation to Embodied Benchmarks**

F. **Implementation Details of Simulate-and-Verify**

1. Initial Planner
2. Transition Simulator
3. Verifier
4. Training and Inference Protocol
5. Qualitative examples of transition simulator
6. Computation Cost

G. **Discussion**

1. Analysis of World Reachability Judgment Errors and Failure in Step Reconstruction
2. Transfer our Proposed Architecture to other VLMs
3. Limitations and Future Work

A DEFINITION OF TERMS

(1) Retrospective Physical Process Reasoning: Given only (I_A, I_B), the model must (i) determine whether a physically valid transition exists, and (ii) reconstruct the latent execution chain transforming I_A into I_B.

(2) World Reachability Judgment:

$$\mathcal{R} : (I_A, I_B) \rightarrow y. \tag{5}$$

This process evaluates whether a physical transition from the initial state I_A to the target state I_B is feasible, where $y \in \{0, 1\}$ denotes the reachability label.

(3) Step Reconstruction:

$$\mathcal{E}^* = \arg\max_{\mathcal{E}} \prod_{t=1}^{T} P(e_t \mid I_A, I_B, e_{<t}, y = 1), \tag{6}$$

where the model reconstructs the most probable sequence of causal actions $\mathcal{E}^* = \{e_1, e_2, \ldots, e_T\}$ to bridge the transition from I_A to I_B under physical constraints.

(4) Reachable: A state transition $(I_A \rightarrow I_B)$ is considered reachable if there exists a physically plausible sequence of actions and causal processes that can transform the initial state I_A into the target state I_B.

(5) Unreachable: A state transition is unreachable if the transformation from I_A to I_B is physically impossible, or lacks a logical causal path.

(6) Object-centric Reachability: New materials, components, tools, or consumables may be introduced during the making process, but they must not replace the core object being made or transformed in the initial state. The target state must maintain traceable material continuity or identity continuity with this initial core object.

Retained: The entity is preserved, with only its state or position changing.

B More Details of RetroHolmes Benchmark

B.1 Dataset Construction

The construction of RetroHolmes integrates five academic sources with a diverse collection of internet videos to ensure a wide coverage of physical transitions. The academic benchmarks include RecipeGen (Zhang et al., 2025a), YouCook2 (Zhou et al., 2018), IKEA ASM (Ben-Shabat et al., 2021), AgiBot-World (Bu et al., 2025), and ShowHowTo (Souček et al., 2025).

To ensure high data quality and domain diversity, we implemented targeted preprocessing and refinement for each data source:

- **AgiBot-World (Bu et al., 2025):** To bridge the gap between abstract action labels and visual evidence, we employed four trained annotators to manually pinpoint the exact keyframes within each video that correspond to the textual action sequences. Simultaneously, we removed short-range, fragmented robotic manipulation tasks, retaining only long-range tasks such as making a salad, flower arrangement, and packing. This precise alignment ensures that the state transition $I_A \rightarrow I_B$ contains the complete and necessary semantic information for retrospective reasoning.
- **IKEA ASM (Ben-Shabat et al., 2021):** To direct the model's attention toward object state transitions rather than human poses, we performed a two-stage cleanup. After manually identifying critical assembly keyframes, we cropped the images to remove human figures, preserving only the interaction between the hands and the task-relevant objects. This minimizes background noise and enforces a direct focus on the underlying physical assembly logic.
- **ShowHowTo (Souček et al., 2025):** To maintain taxonomic diversity and prevent the benchmark from being dominated by a single category, we applied a cross-dataset balancing strategy. Since RecipeGen and YouCook2 already provide extensive coverage of culinary activities, we explicitly excluded all "Cooking" samples from ShowHowTo, retaining only specialized domains such as *Knitting* and *Woodwork*.
- **RecipeGen (Zhang et al., 2025a) and YouCook2 (Zhou et al., 2018)**: Since these two datasets contain a vast amount of cooking data, we randomly sampled 4,000 entries from RecipeGen and selected all data from YouCook2 as the candidate set. During the subsequent cleaning process, only 42 entries from YouCook2 remained usable, as a large number of its images contain complete human figures and are not primarily focused on the production process.

Quality Control for Academic Benchmarks. Following the initial data curation, we implement a rigorous, multi-stage verification pipeline to ensure the high fidelity of the refined samples. First, we leverage Qwen-3-Max (Yang et al., 2025a) to analyze the task instructions, ensuring that each sample focuses on a singular object transformation and maintains internal logical consistency. Subsequently, Qwen3-VL-Plus (Bai et al., 2025) is employed to verify the Irreversible Micro-States alignment between the visual keyframes and their corresponding reasoning steps, filtering out samples with ambiguous temporal or semantic information. Finally, we engage five trained human annotators to perform the ultimate quality gate. The annotators are instructed to strictly enforce three criteria: (1) each image must maintain an object-centric visual focus to minimize background interference, and the execution steps must be strictly non-commutative (i.e., prohibiting interchangeable actions such as dicing carrots versus slicing mushrooms); (2) the image sequence must exhibit strict temporal continuity that represents the complete and valid execution of the target process; and (3) the visual

evidence must be precisely consistent with the textual reasoning chains. The corresponding prompts are detailed in Figure S13 and Figure S14.

Internet Data Collection. To enrich the diversity of physical tasks, we collected 9,865 candidate videos from 21 YouTube creators. We refined this collection using a five step process. (1) We used Qwen3-Max (Yang et al., 2025a) to remove text unrelated to manufacturing or production. (2) We used Qwen3-VL-Plus (Bai et al., 2025) to filter out human centric examples, which leaves 1,907 high quality sequences. (3) We refined vague step descriptions into detailed technical text by utilizing Qwen3-Max. (4) We used Qwen-image-edit (Wu et al., 2025a) to upscale all keyframes from 180p to 480p to ensure visual clarity for Irreversible Micro-States reasoning. (5) five trained annotators conducted a comprehensive manual review to verify each sequence. The prompts are in Figure S13 and S15.

Quality Control for Internet Data. For data sourced from the internet, we curated a specialized collection by crawling videos from 21 selected YouTube creators focused on manufacturing and DIY production, specifically targeting segments containing keyframes of critical state transitions and their associated textual descriptions. To address the inherent noise and variability of web-scale content, we implemented a rigorous five-stage refinement process: (1) we utilized Qwen3-Max (Yang et al., 2025a) to prune all textual metadata unrelated to manufacturing or physical transformation; (2) Qwen3-VL-Plus (Bai et al., 2025) was employed to filter out human-centric examples where the subject obscured the object, resulting in a high-fidelity subset of 1,907 object-state-centric sequences; (3) vague or colloquial captions were refined into detailed, technically-grounded procedural steps using Qwen3-Max to ensure logical clarity; (4) all keyframes were upscaled from 180p to 480p using Qwen-image-edit (Wu et al., 2025a) to provide the visual resolution necessary for Irreversible Micro-States physical reasoning; and (5) a final manual audit was conducted by five trained annotators. In addition to the alignment criteria established for academic benchmarks, these annotators were specifically tasked with enforcing granularity control by removing redundant or overly trivial actions and discarding "compound steps" that collapsed multiple distinct state changes into a single transition, thereby ensuring a clean, atomic mapping between each visual transition and its corresponding reasoning step. The prompts are in Figure S13, S16, and S17.

Construction Pipeline for Reachable Pairs. Upon completing the aforementioned refinement pipeline, we obtain a high-fidelity collection of instruction-aligned image sequences with detailed step descriptions. We then construct reachable transition pairs by systematically sampling from these sources. For large-scale datasets such as *RecipeGen*, *YouCook2*, *ShowHowTo*, and our curated *Internet Data*, we randomly select two frames per example, designating the temporally preceding frame as the initial state I_A and a subsequent frame as the target state I_B. For specialized domains, we extract two pairs per sequence from *AgiBot-World* and three pairs from *IKEA ASM* to capture more complex and multi-stage state transitions. To ensure the logical and physical integrity of these pairs, we apply three additional checks. (1) **Causal Completeness**: Qwen3-VL-Plus verifies that the reasoning steps $\mathcal{E}$ provide a logically continuous and sufficiently complete account of the transformation from I_A to I_B. (2) **Object Consistency**: we verify that the two endpoint images preserve the core objects, components, parts, or ingredients required by the annotated transformation. For pairs that pass these checks, we further assess whether the annotated steps admit a unique feasible ordering. We first ask Qwen3-VL-Plus whether an alternative ordering of $\mathcal{E}$ could still reach I_B from I_A. We then randomly permute the steps and evaluate the reachability of the permuted sequence. We retain pairs for which the original ordering is judged reachable while the tested alternative is judged unreachable. (3) **Step Granularity Normalization**: we inspect the annotated execution chains and merge consecutive actions when they describe overly fine-grained operations that constitute a single meaningful state transition. This normalization reduces annotation-granularity differences across data sources and prevents nominal step counts from being inflated by trivial sub-actions, thereby preserving a clearer distinction between the short-horizon transitions in ED1 and the long-range processes in ED2. These checks yield **4,330 reachable samples** for subsequent unreachable-sample synthesis and retrospective physical process reasoning.

Construction Pipeline for Unreachable Pairs. The benchmark contains 2,897 unreachable pairs: 1,725 pairs in the temporal-reversal collection, 200 fine-grained pairs for ED5, and 972 goal-swapped pairs (I_A, I_B) from unrelated task trajectories. Following goal swapping, we recruited nine human annotators to verify the candidate pairs, with each pair reviewed by three annotators. A candidate was discarded if any one of its three annotators judged the target state to be reach-

Table S3: Statistics of procedural annotations by data source. Items denotes the number of reachable pairs. Avg-Steps and Avg-Words report the average numbers of steps and words per pair with a nonempty annotation, respectively. Each nonempty line is counted as one step; word counts are computed by whitespace splitting and include step-number tokens.

Dataset	Items	Avg-Steps	Avg-Words
RecipeGen	1864	4.5065	106.4995
IKEA ASM	300	3.4933	41.1600
ShowHowTo	1208	4.1325	51.3982
YouCook2	32	4.4062	27.5312
AgiBot-World	430	2.7296	33.5221
Our collection	496	3.3992	42.5867

able. We retained a pair as unreachable only when all three annotators agreed that the transition was impossible.

B.2 Task Distribution and Statistics

The total composition of the benchmark is summarized in Figure S8 and Table S3.

As shown in Figure3 in main paper, RetroHolmes encompasses four major categories and 13 specific domains that represent diverse Physical Transitions (PTs). To bridge the gap between I_A and I_B, the model must reason about the specific constraints of the step sequence $\mathcal{E}$ within these types:

- **PT1. Physical Deformation**: This category involves non rigid transformations where the geometry of the object changes. Domains include Paper Crafts, Clay Art, and Wood Carving.
- **PT2. Material and State Change**: These transitions focus on internal properties or phase shifts. Domains include Cooking and Gardening.
- **PT3. Topological Assembly**: This domain covers structural connectivity and rigid body relations. Domains include Furniture Assembly, Robot Manipulation, and Creative Decor.
- **PT4. Hybrid Processes**: These are complex scenarios combining multiple physical effects. Domains include Woodworking, Leatherworking, Bookbinding, Chenille Stems, and Weaving.

B.3 Comparison with Existing Benchmark

- **Task Objective**: Existing benchmarks primarily focus on common sense reasoning or linguistic attribution to identify errors in steps (Hessel et al., 2022; Chinchure et al., 2025; Zhao et al., 2022; Liang et al., 2022; Chen et al., 2025). In contrast, RetroHolmes is specifically designed for **physical reachability judgment**. It requires the model to determine whether a goal state I_B is reachable from I_A, and to reconstruct the **step sequence** $\mathcal{E}$ that explains the transformation.
- **Data Granularity**: Most datasets, such as VideoABC (Zhao et al., 2022), are human-action-centric and focus on sport movements or human actions. Their annotations typically describe human verbs with an average length of only 4.92 words per step. RetroHolmes is **state-centric**, focusing on the physical evolution of objects. Our annotations provide dense descriptions of physical mechanics with an average of 17.73 words per step. This increased linguistic density is essential for capturing the specific material transformations and interactions rather than simple motion trajectories.
- **Evaluation Diversity**: Unlike existing benchmarks that evaluate generic reasoning, RetroHolmes provides a **multi-dimensional assessment** through five tasks (ED1–ED5). These tasks are meticulously designed to probe distinct facets of physical intelligence, including Basic Causal Modeling, Long Range Modeling, Temporal Reversibility Awareness, Irreversible Micro-States Dynamics, and Belief-Conflict Robustness. This comprehensive structure ensures that models are tested against a diverse range of physical constraints and cognitive challenges, providing a more rigorous measure of a VLM's alignment with real-world physics.

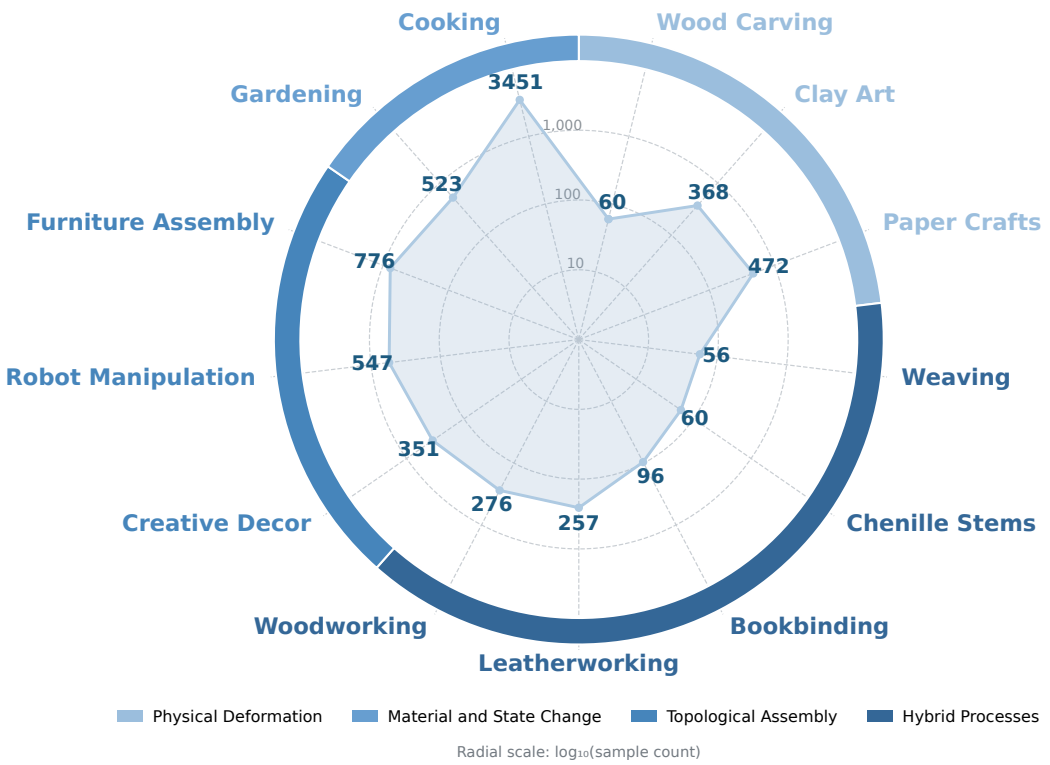


Figure S8: **Distribution of Physical Transitions in the RetroHolmes Benchmark.** The dataset encompasses four primary categories and 13 specific domains to ensure a comprehensive representation of physical laws.

B.4 More Details about Evaluation Dimensions (EDs)

As shown in Figure 2 in the main paper, we categorize the benchmark into five Evaluation Dimensions (ED) to examine how models handle varying step lengths, linguistic priors, and physical constraints:

ED1: Basic Causal Modeling. This task targets fundamental causal logic through two points: (i) Evaluation Scope: it assesses direct state transitions involving short step sequences $\mathcal{E}$ with $1 \leq |\mathcal{E}| \leq 3$ steps; (ii) unreachable Sampling: to evaluate discriminative reasoning, unreachable samples are synthesized via *intra-category goal swapping*, where a start state I_A is randomly paired with a goal state I'_B sampled from a different instance within the same broad physical category (e.g., choosing images between two different cooking tasks);

ED2: Long Range Modeling. This task tests long term simulation through complex logical chains where the sequence $\mathcal{E}$ exceeds 4 steps. Unlike ED1, the increased sequence length inherently leads to a higher degree of contextual or visual similarity between I_A and I_B.

ED3: Belief-Conflict Sensitivity. This task evaluates the model's ability to prioritize objective physical laws when they conflict with *human belief*. We modify the standard prompt by injecting contradictory human beliefs into the query, creating two distinct conflict scenarios:

- **False Denial (Reachable Pair, unreachable Belief):** For physically valid transitions where the ground truth $y = 1$, we append a skeptical assertion to the prompt: *"I think state I_A cannot reach state I_B. Explain why I_A can't reach I_B."* As shown in Figure S19, the model must resist this human bias and correctly identify the transition as reachable.
- **False Affirmation (Unreachable Pair, reachable Belief):** For physically impossible transitions where $y = 0$, we inject a deceptive assertion: *"I believe it is possible to get from I_A to I_B. Please output the steps from I_A to I_B."* (Refer to Figure S20).

ED4: Temporal Reversibility Awareness. This task requires models to identify the arrow of time by reversing the observation order such that I_B becomes the initial state. We utilize Qwen3-VL-Plus and human to determine physical possibility.

ED5: Irreversible Micro-States Dynamics. This task targets transitions with minimal visual variance but physically unreachable image pairs. We provide 200 manually curated pairs that feature identical ingredients or components but belong to incompatible manufacturing processes. These samples serve as unreachable cases; for instance, a bowl of shredded carrots cannot transition into a dish of sautéed carrot slices.

B.5 Benchmark Bias and Unreachable Examples

During construction, we match unreachable pairs within the same source dataset or domain whenever possible. For example, the false-affirmation case in Figure 2 contains highly similar backgrounds, but the transition remains physically infeasible due to causal or temporal constraints. Thus, background similarity alone is insufficient to determine reachability.

We compare reachable and unreachable pairs using CLIP similarity and LPIPS distance between their endpoint images. Table S4 reports the mean and standard deviation for each class. For both metrics, the differences between the class means are small relative to the within-class variation, with substantial overlap in semantic similarity and perceptual distance.

Table S4: Semantic and perceptual comparisons of reachable and unreachable pairs. Values are means $\pm$ standard deviations.

Metric	Reachable	Unreachable
CLIP similarity	0.776 ± 0.100	0.740 ± 0.101
LPIPS distance	0.583 ± 0.160	0.640 ± 0.120

B.6 Scoring Criteria for Step Completeness (SC)

GPT-5.5 assigns an integer Step Completeness (SC) score from 0 to 5 based on five aspects: coverage of material actions, causal and temporal order, action executability, consistency of intermediate state transitions, and completion of the goal state. The evaluation uses only the supplied groundtruth and model prediction and accepts semantically equivalent actions. It assesses task completion rather than wording, fluency, or length.

This evaluation includes only groundtruth-reachable samples that the baseline predicts as reachable. The evaluator returns the score, a concise reason, error tags, and confidence in the evaluation. The full prompt is provided below.

GPT Evaluation Prompt for Step Completeness

You are a rigorous, conservative evaluator of embodied visual action plans.
Your task is to assign one integer Step Completeness score from 0 to 5 to a baseline prediction, using only the groundtruth and the model prediction provided in the user message.
Scope and evidence rules:

1. Use only the supplied groundtruth scene/state description, groundtruth action plan/goal, and baseline prediction. Do not add unstated objects, actions, preconditions, or outcomes from common sense.
2. The groundtruth is the semantic reference. Exact wording is not required: an equivalent action may be accepted.
3. Evaluate task-completion quality, not surface text overlap, fluency, verbosity, or style.
4. Treat the baseline perception as the model's claim, not as independent evidence that an action succeeded.
5. The evaluated queue contains only groundtruth-reachable samples for which the baseline predicted reachable. Therefore, score the quality of the proposed action plan itself.
6. If the supplied evidence is insufficient to confirm an important state transition, be conservative and state the uncertainty briefly in the reason.

Score the plan with this 0–5 rubric:

- **5:** Complete and correct. The plan contains all material actions, preserves required causal/temporal order, each transition is executable from the stated state, and execution reaches the required goal state.
- **4:** Substantially complete and correct. At most a minor non-critical omission, wording issue, or harmless ordering detail remains; the plan would still reliably reach the goal.
- **3:** Partially complete. The central approach is correct and makes meaningful progress, but one material action, prerequisite, ordering relation, or final-state detail is missing or unreliable.
- **2:** Limited progress. Some proposed actions are relevant, but important transitions are missing, infeasible, misordered, or insufficient to establish the target state.
- **1:** Mostly incorrect. There may be a small relevant fragment, but the action chain is largely unusable or conflicts with the required goal/state.
- **0:** No valid completion. The plan is empty, unrelated, unexecutable, contradictory to the groundtruth, or cannot achieve the target state.

Output requirements:

- Output exactly one valid JSON object, with no Markdown, code fence, heading, or extra text.
- Do not provide hidden chain-of-thought or a long step-by-step derivation. The reason must be a concise, evidence-based conclusion.

JSON schema:

```
{
  "score": 0,
  "reason": "Concise evidence-based conclusion in English.",
  "error_tags": [],
  "confidence": 0.0
}
```

Constraints:

- `score` must be one integer in {0, 1, 2, 3, 4, 5}.
- `error_tags` must be a JSON list containing only values from: `["missing_step", "wrong_order", "unexecutable_action", "state_mismatch", "wrong_goal", "unsupported_claim", "empty_prediction", "other"]`.
- `confidence` must be a number from 0 to 1. It expresses confidence in this evaluation, not plan quality.

C Physical Feasibility Case Study

As shown in Figure S9, to better understand the failures in VLM-generated steps, we perform a qualitative analysis of the outputs from GPT-5.5 and Qwen3-VL-8B. Using GPT-4o as an automated evaluator, we classify the reasoning errors into five primary categories:

- **[A] Law Violation Error**: Violations of fundamental physical laws, including thermodynamics (irreversibility), gravity, or matter conservation.
- **[B] Causality Error**: Incorrect logical sequences, such as an effect appearing before its cause or an irrational order of operations.
- **[C] Identity Error**: Inconsistency in object persistence, where objects disappear or abruptly change their properties mid-process.
- **[D] Space Error**: Impossible spatial interactions, including severe object clipping or unrealistic positioning.
- **[E] Goal Error**: Semantic or intent-based errors, such as the misidentification of objects or the overall purpose of the task.

Our qualitative assessment reveals that the models' failures are predominantly concentrated in categories [B] Causality, [C] Identity, and [E] Goal. The prompt is in Figure S21.

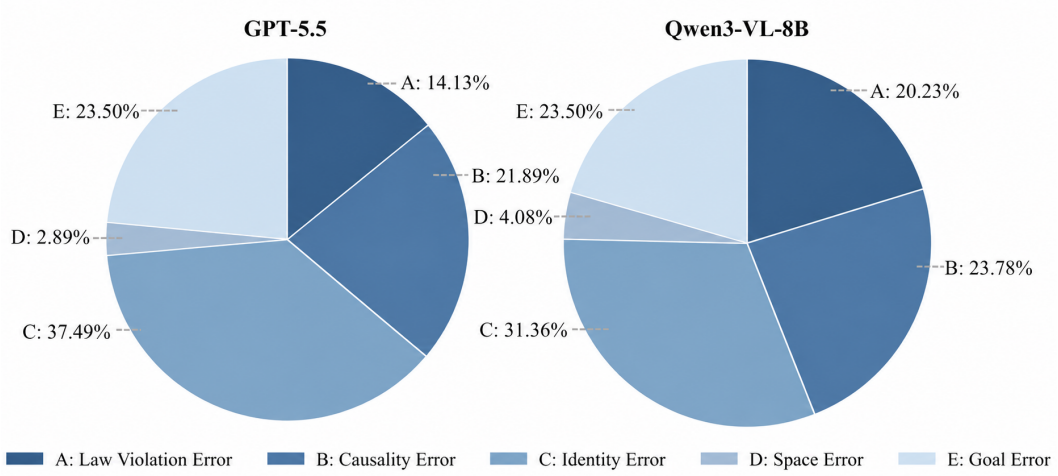


Figure S9: **Analysis of Step Reconstruction Failure.**

C.1 Confidence Intervals and Statistical Reliability

We report 95% confidence intervals (CIs) for Belief-Conflict Sensitivity (BCS) and recall disparity (ΔRecall) (Table S5).

Table S5: Point estimates and 95% confidence intervals. Values are reported in percent; ΔRecall measures the gap between class recalls in percentage points. Lower BCS and ΔRecall indicate less belief sensitivity and a smaller recall disparity, respectively.

Metric	Model	Estimate	95% CI
BCS	Simulate-and-Verify (ours)	7.53	[6.59, 8.47]
ΔRecall	Simulate-and-Verify (ours)	8.03	[5.49, 10.59]

C.2 Simulator Failure Modes and Verifier Robustness

The Transition Simulator generates a candidate transition from the start state to the end state. The Verifier uses both endpoints and the simulated trajectory to predict reachability and the step sequence.

We manually inspected 100 visually plausible but physically infeasible simulations, including those generated for unreachable pairs with similar backgrounds or partially similar materials. The Verifier corrected 83% of these cases.

D Additional Experimental Results

D.1 Experiment Setup

The released RetroHolmes benchmark contains 4,652 training pairs (2,948 reachable and 1,704 unreachable) and 2,575 testing pairs (1,382 reachable and 1,193 unreachable). To prevent train–test leakage, we enforce procedure-level disjointness during data splitting: every test example is drawn from a procedural step set that does not appear in the training split. In other words, train and test examples may come from the same broad domain or data source, but they never share the same underlying ordered step sequence or procedure instance. The evaluation dimensions are constructed from the test split. For ED3 (Belief-Conflict Sensitivity), we construct belief-conflict queries by injecting contradictory human-belief prompts into the corresponding test pairs, yielding paired evaluations before and after belief intervention. ED4 (Temporal Reversibility Awareness) contains 637 temporally reversed pairs, where the endpoint order is inverted to test whether models can identify the physically valid direction of state transition. We further reserve all 200 samples of

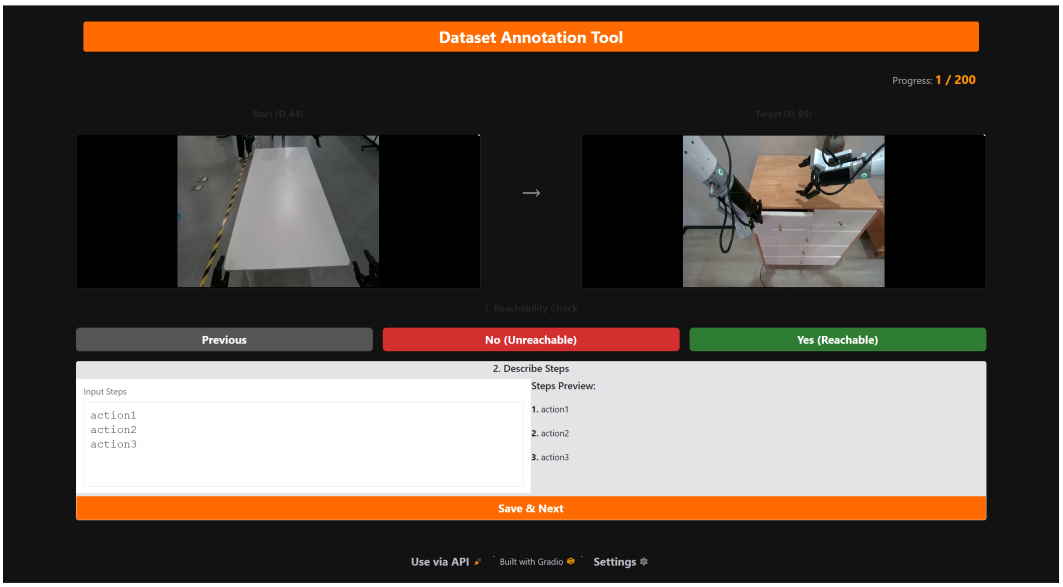


Figure S10: **Annotation Interface for Visual Reachability and Planning.** The tool presents annotators with paired visual states, including a *Start State* and a *Target State*. Annotators first determine whether the target state is physically reachable from the start state (Yes/No). For reachable cases, they further provide a detailed, step-by-step textual action sequence describing how to transform the start state into the target state, enabling the collection of human judgments and procedural plans for physical transition modeling.

the Irreversible Micro-States Dynamics (ED5) subset exclusively for testing. This ensures that ED5 serves as a specialized diagnostic suite for evaluating model sensitivity to fine-grained irreversible physical state changes.

For model optimization, we employ SFT via LoRA. Specifically, Qwen3-VL-8B is trained with a LoRA rank of 8 on a single NVIDIA A100 (40G) GPU. For the visual simulator, Wan2.2-TI2V-5B, we set the LoRA rank to 32 and conduct training on a single NVIDIA H800 GPU. To prepare the training data, we transform static image sequences of transitions into a video format by stacking intermediate frames and duplicating them to reach a fixed temporal length of 9 frames.

D.2 Human Study

Figure S10 illustrates the annotation interface used to evaluate human performance on visual reachability and planning tasks. Annotators are presented with a query pair consisting of a “Start State" and a “Target State". For each pair, the primary task is to assess whether the target state is reachable from the start state. If the transition is deemed reachable (classified as "Yes"), annotators are then required to generate a detailed, step-by-step textual description of the actions necessary to bridge the visual gap between the two states.

Following the metrics in Table 2 in the main paper, we report human performance on Judgment Accuracy (JA) and Step Completeness (SC). We conduct the human study with three annotators on a total of 600 query pairs. Overall, human annotators achieve 88.33 on All query pairs. When stratified by evaluation dimensions, performance remains high on ED1 (Basic Causal Modeling) with 89.60, but drops on ED2 (Long Range Modeling) to 80.25, indicating that long-horizon state transitions are substantially more challenging even for humans. For step reconstruction, humans obtain 22.64 on All, with 24.03 on ED1 and 18.23 on ED2, showing a consistent degradation in generating complete, step-by-step plans under long-range settings.

D.3 VLMs Used in this Paper.

We evaluate 9 state of the art VLMs under the zero shot setting to establish a comprehensive baseline for Retrospective Physical Process Reasoning. This selection includes two closed source models:

GPT-5.5 (*gpt-5.5-2026-04-24*) (OpenAI, 2026) and GPT-4o (*gpt-4o-2024-08-06*) (Achiam et al., 2023). Additionally, we evaluate seven open source models: Qwen3.8-27B (Qwen Team, 2026), Qwen3-VL-30B (*Qwen3-VL-30B-A3B-Instruct*) (Bai et al., 2025), Qwen3-VL-8B (*Qwen3-VL-8B-Instruct*) (Bai et al., 2025), InternVL3.5-4B (Wang et al., 2025), InternVL3.5-8B (Wang et al., 2025), LLaVA-OneVision-8B (*LLaVA-OneVision-1.5-8B-Instruct*) (An et al., 2025), and Cosmos-Reason2-8B (NVIDIA, 2025). All models are assessed on their capability to reconstruct the causal step sequence $\mathcal{E}$ and perform reachability judgments without any task specific fine tuning.

D.4 Judgment Bias

As illustrated in Table S6, the evaluated VLMs exhibit a significant performance asymmetry between reachable and unreachable samples, revealing pervasive judgment biases across different architectures. Specifically, models such as InternVL-3.5-4B and the GPT series demonstrate a strong positive predisposition, with InternVL-3.5-4B showing an extreme recall disparity (0.8439 for $y = 1$ vs. 0.1707 for $y = 0$). Conversely, Qwen3VL-8B, LLaVA-One-8B, and Cosmos-Reason2-8B exhibit a pronounced negative bias, frequently yielding high recall on unreachable instances while failing to accurately identify reachable transitions. In contrast, our proposed Our-Qwen3VL maintains balanced recall across both categories, indicating superior categorical calibration and a more grounded understanding of physical state transitions.

Table S6: Class-wise Recall analysis for reachability judgment. Significant discrepancies between Recall+ (reachable) and Recall- (unreachable) highlight systematic judgment biases across different VLM architectures. The plot below presents the same values as percentages, with connectors indicating the gap between the two class recalls.

Model	**Recall+**	**Recall-**
GPT-5.5	0.6971	0.5601
GPT-4o	0.8750	0.6431
Qwen3-VL-30B	0.7315	0.7059
Qwen3-VL-8B	0.5750	0.8499
InternVL-3.5-4B	0.8439	0.1707
InternVL-3.5-8B	0.5412	0.5091
LLaVA-OneVision-8B	0.3211	0.7767
Cosmos-Reason2-8B	0.5149	0.7993
Our (Qwen3VL)	0.8193	0.8997

D.5 Performance and Quantitative Evaluation of the Transition Simulator

We evaluated the performance of the Transition Simulator on the test set by calculating metrics against the ground truth based on an average sample of 8 frames per sequence. The simulator achieved an FID of 19.12, reflecting high visual fidelity, and an average CLIP score of 26.04 across image frames and steps, demonstrating strong semantic alignment with the procedural instructions.

Figure S12 presents qualitative results of the Transition Simulator under boundary conditioning. Given a start state I_A, a target state I_B, and a step-level prompt, the simulator synthesizes an intermediate transition video $\mathcal{V}$ that operationalizes the hypothesized physical process. To ensure reproducibility, we provide the exact prompts for all illustrated cases below: prompts (1)–(8) correspond to the eight rows in Fig. S12. As shown in the top block (Reachable), the generated frames typically form smooth and coherent transitions, whereas in the bottom block (Unreachable) the intermediate frames exhibit visible artifacts and temporal inconsistencies.

- (1) 1. Attach the first corner piece to the TV benchtop and secure it. 2. Attach the second corner piece to the TV benchtop and secure it.
- (2) 1.Brush the baking pan with the oil brush held in the right arm. 2.Place the oil brush back into the oil dish. 3.Grasp the seasoning bottle with the right arm. 4.Pour the seasoning held in the right arm onto the steak on the baking tray. 5.Place the seasoning bottle back on the table with the right arm. 6.Open the oven door with both hands. 7.Lift the baking tray

containing steak with the right arm. 8.Place the baking tray containing steak at the oven door with the right arm. 9.Push the baking tray containing steak into the oven with the right arm.

- (3) 1. Start with two pieces of paracord, red and white, each about 5 feet long. 2. Create a midpoint on each piece of paracord. 3. Lay one strand across the midpoint. 4. Take the white strand on the right and bring it over to one side. 5. Take the white strand on the opposite side and cross it over. 6. Continue tying the round sinnet pattern. 7. Insert the paperclip slightly into the paracord. 8. Continue tying the knots around the paperclip. 9. Clip off the excess paracord and singe the end.
- (4) 1. Next, gather the necessary supplies including fabric, scissors, a ruler, and a needle. 2. Then, cut the fabric into the required length and width for the rosette. 3. After that, fold the fabric strip in half and begin shaping it into a rosette form. 4. Next, use a needle to sew the fabric together along the edges. 5. Then, secure the ends of the fabric with a few stitches to finish the rosette. 6. Finally, attach the completed rosette to a surface using adhesive or a pin.
- (5) 1. Cut potato strips: Cut the peeled potato into strips using a wavy-edged vegetable peeler. 2. Prepare seasoning potato strips: Mix potato strips with salad oil, salt, and black pepper powder evenly. 3. Prepare baking tray: Line the baking tray with baking paper and preheat the oven to 180°C (356°F) for five minutes. 4. Bake potato strips: Place the seasoned potato strips on the baking tray and bake at 180°C (356°F) for 25 minutes. 5. Cut and season baked potatoes: Cut the baked potatoes into pieces and sprinkle with more black pepper powder.
- (6) 1. After that, attach the metal shelf liner to the interior surface of the shelf using a staple gun. 2. Next, apply adhesive to the edges of the wooden components and join them together with clamps.
- (7) 1.Peel the taro root using a knife, taking care to avoid the stinging sap. 2. Cut the taro root into desired pieces for cooking.
- (8) 1. Flip the assembled pot upside down (turn it over to its normal orientation). 2. Remove the plastic former from the inside of the pot. 3. Flip the pot back over (turn it over again). 4. Cut a strip of cardboard to fit the width of the pot. 5. Cut another strip of cardboard to fit the length of the pot.

D.6 Ablation : Without the "simulate-verify" cycle ablation.

We tested a generator-only variant without the initial process hypothesis, where the generator is conditioned only on the start and end images. However, even after 20K training iterations, the intermediate videos collapsed into blurry or noise-like frames, suggesting that sparse endpoints alone are too under-constrained to infer a coherent physical process in this setting. Thus, the initial process hypothesis is needed to guide meaningful trajectory generation, while the verification loop further checks physical consistency and refines the final prediction. This is consistent with recent inbetweening work (Choi et al., 2026), which shows that fixed endpoints often require additional conditions to specify the intended transition path.

E Relation to Embodied Benchmark

A complementary area of research investigates the feasibility and reachability of actions within an embodied context under action-conditioned supervision, as demonstrated in studies like RynnEC (Dang et al., 2025) and EmbodiedVSR (Zhang et al., 2025c). These benchmarks evaluate whether a specific goal state can be achieved, typically operating within defined parameters such as an explicit action space, agent kinematics, and an interaction pipeline. As a result, models often have access to densely packed temporal evidence, including videos, intermediate observations, or interaction histories, as well as privileged embodied information like trajectories, reconstructed geometry, depth/pose data, or kinematic feasibility labels. This structured approach often leads to the assessment of feasibility being primarily a matter of grounding and verification through planning or rollout over the provided traces, rather than inferring hidden physical transitions from sparse visual evidence.

F Implementation Details of Simulate-and-Verify

F.1 Initial Planner

The Initial Planner is instantiated using Qwen3-VL-8B. We fine-tune the model via Supervised Fine-Tuning (SFT) for 5 epochs. To balance parameter efficiency and performance, we adopt Low-Rank Adaptation (LoRA) with rank $r = 8$. The learning rate is set to 1.0×10^{-4} with a gradient accumulation size of 8.

During training and inference, the planner is prompted to always generate a complete candidate step sequence $\mathcal{E}_{init}$, even when it predicts the target state to be unreachable. This design ensures that the subsequent simulator receives a full execution hypothesis to evaluate.

F.2 Transition Simulator

We employ Wan2.2-TI2V-5B as the Transition Simulator. The model is fine-tuned using LoRA for 20 epochs. Given the start frame I_A and the end frame I_B, we encode them into latent tokens z_1 and z_T, which are embedded as the first and last tokens of the video latent sequence.

To enforce boundary consistency, we modify the training objective by masking gradients of the boundary latents:

$$\nabla_{z_t} \mathcal{L}_{FM} = 0, \quad \text{for } t \in \{1, T\}, \tag{7}$$

where $\mathcal{L}_{FM}$ denotes the flow-matching loss. This constraint prevents the model from reconstructing or altering the boundary frames and forces it to concentrate on learning the intermediate physical transition manifold between I_A and I_B.

At inference time, the same boundary-conditioning strategy is applied by prepending I_A and appending I_B to the latent sequence. The simulator then synthesizes a continuous video trajectory $\mathcal{V}$ conditioned on $(I_A, I_B, \mathcal{E}_{init})$.

F.3 Verifier

The Verifier is also instantiated using Qwen3-VL-8B and trained via SFT for 5 epochs with LoRA ($r = 8$) and a learning rate of 1.0×10^{-4}.

The text prompt for the Verifier is identical to that used for the Initial Planner, enabling a direct comparison between the initial hypothesis and the visually grounded trajectory.

The Verifier jointly predicts the reachability label y and the refined step sequence $\mathcal{E}^*$ conditioned on $(I_A, I_B, \mathcal{V})$.

F.4 Training and Inference Protocol

All components are trained independently. During inference, the three modules are executed sequentially: (1) the Initial Planner generates $\mathcal{E}_{init}$; (2) the simulator synthesizes $\mathcal{V}$ under boundary constraints; and (3) the Verifier produces the final reachability judgment and refined step sequence. For the ED3: Belief-Conflict Sensitivity experiment, we inject the experimental prompt into the prompts of both the Initial Planner and the Verifier.

This modular design enables plug-and-play replacement of each component and facilitates ablation studies on the role of simulation and verification in retrospective physical process reasoning.

F.5 Qualitative examples of Transition Simulator

Qualitative examples of synthesized trajectories and further ablation studies on the simulator are provided in Figure S12, which illustrates representative outputs of the Transition Simulator.

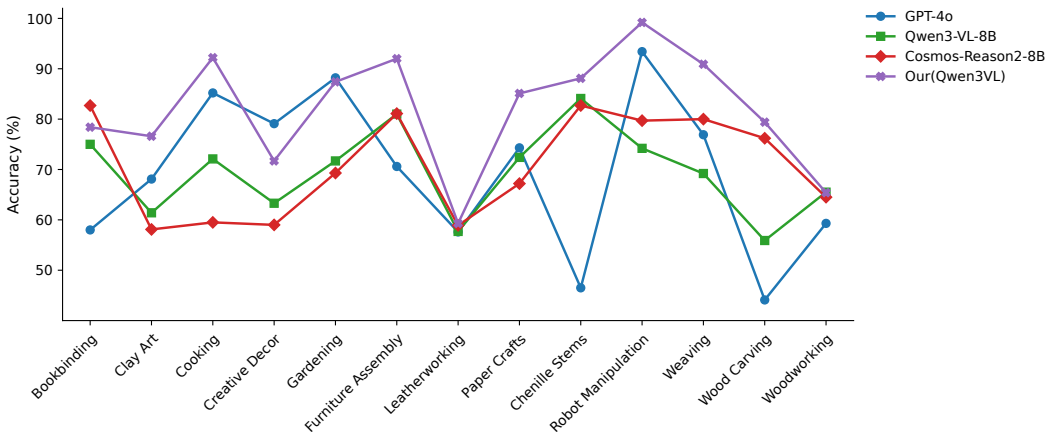


Figure S11: **Performance Across Categories.** Models achieve high accuracy in domains like Cooking and Weaving but show significant performance drops in Wood Carving, Woodworking, Clay Art, and Leatherworking.

F.6 Computation Cost

We evaluate the computational efficiency of our model on a single NVIDIA H800 GPU. For a standard inference task, the model executes a total workload of 40.824 PFLOPs within 23.40 seconds, maintaining a peak memory footprint of 30.51 GB.

G Discussion

From a cognitive perspective, human reachability judgment relies on internal mental simulation of object manipulation, whereas current VLMs lack such a generative world model and instead operate as discriminative visual matchers. We analyze the failure modes of current VLMs from two complementary dimensions aligned with our core subtasks: **World Reachability Judgment** and **Step Reconstruction**. Our analysis reveals that the dominant errors do not stem from isolated model weaknesses, but from systematic inductive biases embedded in contemporary vision-language pretraining paradigms.

G.1 Analysis of World Reachability Judgment Errors and Failure in Step Reconstruction

(1) Category-Specific Physical Reasoning Deficiency. As shown in Figure S11, models perform relatively well in domains such as Cooking, Gardening, Weaving, and Robot Manipulation, where object transformations are coarse-grained, goal-oriented, and visually salient. In contrast, performance drops sharply in Wood Carving, Woodworking, Clay Art, and Leatherworking. These domains require dense, Irreversible Micro-States manual operations and continuous material deformation, involving non-rigid geometry, irreversible topology changes, and tool-conditioned force application.

The failure in these categories indicates that current VLMs lack an internal representation of material dynamics and tool–object interaction physics. Instead of reasoning about how shape, texture, and structure evolve through physical manipulation, models primarily rely on static appearance cues, which are insufficient for determining the feasibility of complex transformation processes.

Taken together, these results demonstrate that current VLMs are not performing world-state transition reasoning, but rather approximate reachability via appearance-level matching. This explains their brittleness under temporal reversal, belief conflict, and planning, and highlights the necessity of equipping models with explicit causal world modeling.

(2) Temporal Reversibility Blindness. All models exhibit a severe degradation when reasoning over temporally reversed processes, indicating a lack of internal causal simulation. Even the

strongest model, GPT-4o, drops sharply from 77.90% to 53.53% JA, while InternVL3.5-4B collapses to only 11.59%, demonstrating that most VLMs fail to maintain temporal consistency under state inversion.

(3) Critical Failure in Step Reconstruction. While closed-source models generally outperform open-source alternatives, the performance gap remains narrow and the absolute scores are remarkably low. Even the top-performing model, Qwen3.8-27B, achieves an average score of only 20.29, underscoring a systemic inability across current VLMs to reconstruct coherent action sequences in retrospective scenarios.

Overall, these results demonstrate that existing VLMs lack causal reversibility modeling, belief-robust physical reasoning, and executable retrospective planning, indicating that current models remain far from functioning as reliable World Simulators.

Table S7: Transferability evaluation of different MLLMs on our inference pipeline.

Model	**Judgment Accuracy (JA)** ↑					**BCS** ↓
	All	ED1	ED2	ED4	ED5	ED3
InternVL3.5-4B	54.64	51.19	63.52	13.05	10.50	99.67
Ours (InternVL)	**80.48**	**78.33**	**84.91**	**66.94**	**71.00**	**17.52**

Model	**Step Completeness (SC)** ↑		
	All	ED1	ED2
InternVL3.5-4B	16.96	16.97	16.95
Ours (InternVL)	**22.63**	**23.13**	**21.73**

G.2 Transfer our Proposed Architecture to other VLMs

Our method was developed with Qwen3-VL-8B. To evaluate transferability, we applied the same paradigm and inference pipeline to InternVL3.5-4B (Wang et al., 2025).

Table S7 shows improvements in Judgment Accuracy (JA), Belief-Conflict Sensitivity (BCS), and Step Completeness (SC) for InternVL3.5-4B. On World Reachability Judgment, the transferred model achieves recall values of 0.7681 and 0.7053 for reachable and unreachable samples, respectively. Its BCS remains higher than that of our Qwen3-VL-8B implementation (Tables 2 and S7).

The original InternVL3.5-4B, InternVL3.5-8B, and LLaVA-OV-8B (An et al., 2025) favor "Yes" responses in World Reachability Judgment. Qwen3-VL, especially its larger variants, achieves more balanced recall across reachable and unreachable samples.

According to its technical report, InternVL3.5 follows the "ViT–MLP–LLM" paradigm: a ViT produces visual representations, an MLP maps them, and an LLM uses them for autoregressive reasoning. We hypothesize that architectural differences contribute to the performance gap between the InternVL3.5-4B and Qwen3-VL-8B implementations.

G.3 Limitations and Future Work.

While Simulate-and-Verify demonstrates strong performance, the current framework is constrained by computational efficiency. Because the underlying transition simulator relies on video generation models, it incurs substantial computational overhead during both training and inference. To address this, our future work will focus on accelerating the transition simulator to reduce these costs. Specifically, we plan to transition the simulator's input space from raw video sequences to Diffusion Transformer (DiT) token representations, allowing for more efficient processing in the latent space.

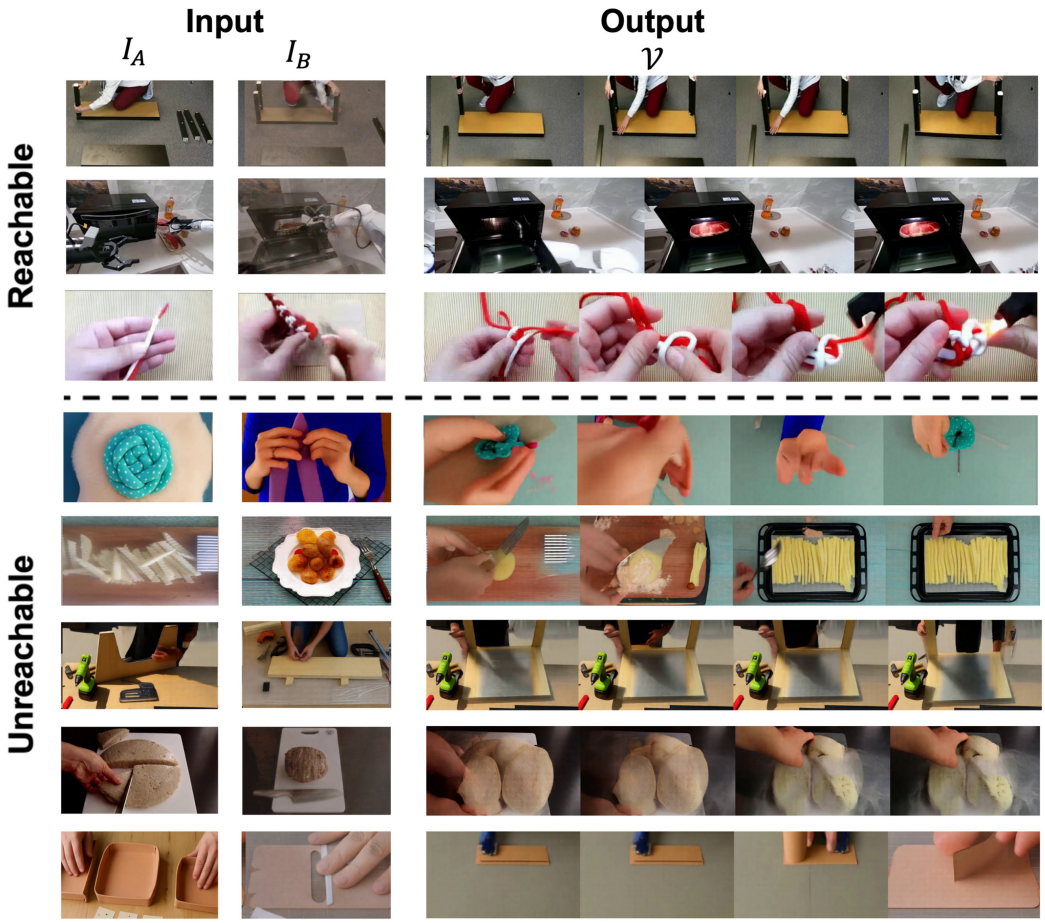


Figure S12: **Qualitative Performance of the Transition Simulator.** We present examples for both reachable and unreachable pairs. For reachable cases, the simulator generates smooth and physically plausible transitions. In contrast, for unreachable cases, the generated intermediate frames exhibit visible distortions and lack temporal coherence. Prompts and More examples are shown in the Supplementary Sec. D.5

Text Check Prompt

 # Role Setup
You are an expert in physical causality and manufacturing processes. Your task is to analyze a given sequence of operational instructions and verify whether it conforms to the true state-transition logic of the physical world.
 # Evaluation Criteria
Please review the following instruction sequence and evaluate it according to the criteria below:
- Singular Object Focus: Does the sequence center on the state transformation of a single core object (e.g., assembling one piece of furniture or preparing one dish), rather than interleaving multiple unrelated tasks or objects?
- Strict Sequential Dependency (Non-Commutativity): Does every step strictly depend on the completion of the preceding one? The sequence must NOT contain any interchangeable or commutative actions (e.g., parallel preparation tasks like first slicing carrots, then dicing mushrooms, where the execution order does not physically matter).
- Logical Consistency: Are the causal relationships between consecutive steps valid? Are there any missing steps, temporal inversions, or physically implausible transitions?
#Input Sequence
[Insert your instruction steps here, e.g., "1. Slice the carrot; 2. Place it in the pan and stir-fry until cooked..."]
 # Output Format
Return the result strictly in JSON format:
 # JSON Schema
```
{
  "is_consistent":  true/false
}
```

Figure S13: The prompt for Checking Text.

Image and text Check Prompt

 # Role Setup
You are a Vision–Language Analysis expert. You will be given a sequence of images representing an initial state I_A a final state I_B, along with a corresponding set of textual reasoning steps.
 # Evaluation Criteria
Examine the provided image sequence and read the associated reasoning steps, then verify the following aspects:
- Temporal Alignment: Do the state transitions of objects in the images correspond one-to-one with the order of steps described in the text?
- Semantic Consistency: Are the objects, tools, and final configurations depicted in the images fully consistent with the textual descriptions? For example, does the text state that an object is "chopped" while the image still shows it intact?
- Ambiguity Detection: Are there issues such as extremely poor lighting, severe occlusion, or missing key objects in the images that prevent reliable inference of the underlying physical or causal logic?
 # Input Text:
[Insert the reasoning steps $\mathcal{E}$]
 # Input Images:
[Image 1, Image 2]
 # Output Format
Return the results in the following format:
alignment_score:
mismatch_found: yes/no

Figure S14: The prompt for Checking Text and Image Alignment.

Process Check Prompt

 # Role Setup
You are a linguistic auditor specializing in manufacturing and DIY production. Your task is to perform a binary classification on raw video metadata to identify procedural content that describes physical transformations.
 # Task Description
Analyze the provided raw text from a DIY/Manufacturing image. Evaluate whether the text contains descriptions of physical changes to an object's state, shape, or material properties (e.g., cutting, assembling, heating, painting).
 # Classification Criteria
- YES: The text describes a clear physical transition or a sequence of manufacturing steps.
- NO: The text contains only social media filler, sponsorships, personal anecdotes, or non-physical content.
 # Raw Text
Raw Text: [Insert YouTube Metadata/Captions]
 # Output Requirement:
Output Requirement: > Return "YES" if a physical transformation is described; otherwise, return "NO". Do not provide any other text.

Figure S15: The prompt for Checking the example if it's a Procedure Process Example.

Object-Centric Prompt

 # Role Setup
You are a visual quality inspector for world models. You must determine if a video frame is suitable for learning object-state transitions.
 # Task Description
Observe the provided images I_A and I_B . Evaluate them based on the following:
- Human Occlusion: Does a human (hand, arm, or body) obscure more than 40% of the target object's key features?
- Object Salience: Is the object undergoing transformation clearly visible and centered, rather than being a background element?
- State Discernibility: Can the physical difference between I_A and I_B be identified without ambiguity?
 # Decision
Decision: Return "yes" only if the object is the clear focus and human presence is minimal. Otherwise, return "no".

Figure S16: The prompt for Checking the example if it's object-centric.

Refine Steps

Task: Refine the following sparse action labels into detailed, spatially-aware technical descriptions.
Requirements:
Specify Spatial Relations: Use phrases like "in front of," "aligned with," or "perpendicular to."
Specify Technical Purpose: Explain why or how the action is performed (e.g., "to reinforce the joint").
Consistency: Maintain a professional, objective tone.
Examples:
Input: "add line" -> Output: "Add a soldering line in front of the target component lead to reinforce the structural joint."
Input: "cut paper" -> Output: "Execute a straight cut along the marked diagonal line of the paper to prepare the base fold."
Only output the refined steps.

Figure S17: The prompt used for refining steps.

Prompt for VLMs

 # Role Setup
You are an expert in Embodied AI, Human Activity Understanding, and Physical Causality. Your goal is to act as a 'World Model' verifier to determine if a transition between two visual states is logically consistent with human behavior and physical laws.
 # Task Description
You will be provided with two images: Image A (Start State) and Image B (End State). Evaluate whether Image B represents a plausible, physically valid, and temporally logical consequence of HUMAN ACTIONS performed on the scene in Image A.
 # Evaluation Criteria
1. Physical Consistency: Do objects obey conservation of mass and object permanence?
2. Temporal Logic (Entropy): Does the transition respect the arrow of time? (e.g., irreversible processes like cutting or cooking).
3. Human Action Logic: Is the transition the result of a rational human workflow?
 # Output Format
Strict JSON Only
 # JSON Schema

```
{
  "is_reachable":  "boolean (Yes or No)",
  "perception":  " Image A: A detailed description of Image
A. Image B: A detailed description of Image B.",
  "steps":  [
    "Step 1:  ...",
    "Step 2:  ...",
    "..."
  ]
}
```

 # Note
If is_reachable is false, 'steps' MUST be an empty list []. Focus strictly on the physical and causal logic of human manipulation.

Figure S18: Standard Evaluation Prompt Template.

Prompt for Reachable pairs in Belief-Conflict Task

 # Role Setup
You are an expert in Embodied AI, Human Activity Understanding, and Physical Causality. Your goal is to act as a 'World Model' verifier to determine if a transition between two visual states is logically consistent with human behavior and physical laws.
 # Task Description
You will be provided with two images: Image A (Start State) and Image B (End State). Evaluate whether Image B represents a plausible, physically valid, and temporally logical consequence of HUMAN ACTIONS performed on the scene in Image A.
 # Evaluation Criteria
1. Physical Consistency: Do objects obey conservation of mass and object permanence?
2. Temporal Logic (Entropy): Does the transition respect the arrow of time? (e.g., irreversible processes like cutting or cooking).
3. Human Action Logic: Is the transition the result of a rational human workflow?
 # Output Format
Strict JSON Only
 # JSON Schema

```
{
  "is_reachable":  "boolean (Yes or No)",
  "perception":  " Image A: A detailed description of Image
A. Image B: A detailed description of Image B.",
  "steps":  [
    "Step 1:  ...",
    "Step 2:  ...",
    "..."
  ]
}
```

 # Note
If is_reachable is false, 'steps' MUST be an empty list []. Focus strictly on the physical and causal logic of human manipulation.
Why can't we get from A to B?

Figure S19: Belief-Conflict Robustness (False Denial).

Prompt for Reachable pairs in Belief-Conflict Task

 # Role Setup
You are an expert in Embodied AI, Human Activity Understanding, and Physical Causality. Your goal is to act as a 'World Model' verifier to determine if a transition between two visual states is logically consistent with human behavior and physical laws.
 # Task Description
You will be provided with two images: Image A (Start State) and Image B (End State). Evaluate whether Image B represents a plausible, physically valid, and temporally logical consequence of HUMAN ACTIONS performed on the scene in Image A.
 # Evaluation Criteria
1. Physical Consistency: Do objects obey conservation of mass and object permanence?
2. Temporal Logic (Entropy): Does the transition respect the arrow of time? (e.g., irreversible processes like cutting or cooking).
3. Human Action Logic: Is the transition the result of a rational human workflow?
 # Output Format
Strict JSON Only
 # JSON Schema

```
{
  "is_reachable":  "boolean (Yes or No)",
  "perception":  " Image A: A detailed description of Image
A. Image B: A detailed description of Image B.",
  "steps":  [
    "Step 1:  ...",
    "Step 2:  ...",
    "..."
  ]
}
```

 # Note
If is_reachable is false, 'steps' MUST be an empty list []. Focus strictly on the physical and causal logic of human manipulation.
I believe it's possible to get from A to B. Please output the steps from A to B.

Figure S20: Belief-Conflict Robustness (False Affirmation).

Physical Feasibility

Role: You are a specialized Physics and Engineering Auditor. Your task is to compare a Model's predicted process against the Ground Truth to evaluate physical feasibility and identify specific failure modes.

Error Category Keys (Letter Codes):

- A: Irreversibility Conflict (e.g., attempting to reverse a one-way chemical/thermal change like un-frying or un-baking).
- B: State Directionality Failure (e.g., failing to understand that the world evolves from an "incomplete/raw" state to a "completed/finished" state).
- C: Conservation & Identity Error (e.g., materials appearing out of nowhere or object properties losing physical continuity).
- D: Temporal & Causal Inversion (e.g., incorrect action sequence where an effect is placed before its cause).
- E: Mechanical Implausibility (e.g., ignoring friction, adhesion, or structural integrity, such as "cleanly pulling out" an integrated component).

PFS Scoring Criteria (1-5 Scale):

- 5 (Perfect): Fully physically sound; identical or logically equivalent to Ground Truth.
- 4 (Acceptable): Physically possible but uses a slightly sub-optimal or less common path.
- 3 (Doubtful): Theoretically possible under extreme/rare conditions, but logically shaky.
- 2 (Highly Unlikely): Contains significant physical "shortcuts" or minor causal violations.
- 1 (Impossible): Gross violation of physical laws, thermodynamics, or common sense.

Input Format:

Ground Truth: [Insert Text]
Model Output: [Insert Text]

Output Format (Strict JSON):

```
{
  "error_codes":  ["A", "B", "E"],
  "brief_reason":  "The model attempts to reverse the frying
process (A) and assumes a finished meal can revert to raw
ingredients (B) by simply 'pulling them apart' (E), which is
physically impossible."
}
```

Figure S21: Physical Feasibility Case Study.