ChronoVision: Temporal Reasoning via Latent State Reconstruction

Yifan Shen[1,2] Jian Xu[1] Boyi Li[1] Yuner Zhang[3] Tianjiao Yu[1]
Bingxuan Li[1] Houze Yang[1] Rushi Wang[1] Xu Cao[1,2]
[1]University of Illinois Urbana-Champaign [2]PediaMed AI [3]University of Pennsylvania
{yifan26,xucao2}@illinois.edu

Abstract

Multimodal large language models excel at passive perception but struggle with complex visual cognitive tasks requiring multi-step temporal reasoning. This degradation largely stems from the inherent ambiguity of language-based reasoning, which often fails to accurately articulate continuous visual transformations. To address this, we propose ChronoVision, a multimodal framework designed to align visual logic with latent imagery. During supervised fine-tuning, a Reconstructive Visual Head predicts the latent representation of the final transformed state, while an ROI Attention Locating module focuses the model on key visual evidence via semantic span queries. In post-training, we apply reinforcement learning with an implicit process grounding mechanism, guided by a composite reward function that evaluates outcome correctness, latent process alignment, and unsupervised visual focus. Furthermore, we introduce Vbvr-VQA, a novel dataset that evaluates temporal tracking by reformulating video reasoning into a strict image-ordering task. Experiments demonstrate that ChronoVision achieves state-of-the-art performance on Vbvr-VQA with 74.8% in-domain and 71.6% out-of-domain accuracy, alongside a strong 55.0% accuracy on IntPhys2, a highly challenging cross-domain benchmark. https://pediamedai.com/Cognition-MLLM/ChronoVision/

1 Introduction

Multimodal Large Language Models have achieved remarkable success in open vocabulary perception tasks such as image captioning and object detection (Liu et al., 2023). However, their capabilities remain highly limited when faced with complex cognitive tasks that require spatial intelligence (Zhang et al., 2026d) and visual reasoning (Li et al., 2024; Schulze Buschoff et al., 2025). In contrast, human cognition intuitively bridges this gap through mental simulation. When observing physical events, such as an object rotating in space, humans can effortlessly project spatial trajectories in brain and predict continuous visual transformations. This cognitive ability to internally simulate and predict the evolution of visual states is known as visual imagery. Current models struggle with tasks that demand visual memory and multiple-step temporal reasoning, in which they rely on recognizing static visual content rather than understanding continuous space with real physical properties.

Current mainstream approaches attempt to enhance reasoning capabilities by extending Chain of Thought reasoning within the textual space (Wei et al., 2022; Zhang et al., 2023b). This creates a fundamental bottleneck. Natural language cannot precisely articulate continuous visual transformations. Attempting to verbally describe a complex three-dimensional rotation or a continuous physical motion trajectory inevitably leads to the loss of critical spatial information (Schulze Buschoff et al., 2025). Therefore, purely text-based reasoning is inadequate for solving cognitive tasks that rely heavily on visual imagery (Qin et al., 2025).

Deeper analysis also reveals significant flaws in existing evaluation methods (Cao et al., 2024; Khezresmaeilzadeh et al., 2026). Traditional Visual Question Answering tasks typically adopt a multiple-choice format. This design introduces a critical issue. The textual descriptions in the question stem or options can be ambiguous, allowing models to exploit linguistic pattern-matching rather than performing authentic visual reasoning. A model might guess the correct option based on text semantics without referencing the visual observation of the transformed result. Furthermore, selecting a single final answer fails to evaluate temporal sequence reasoning capabilities. Genuine visual reasoning requires tracking the evolutionary trajectory from an initial state, through intermediate states, to a final outcome.

To address these limitations, we propose a rig-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 18374–18405
October 24-29, 2026 ©2026 Association for Computational Linguistics

orous new evaluation paradigm and a novel multimodal reasoning framework. First, we present Vbvr-VQA, a benchmark that reformulates video reasoning as a strict image-ordering task (Wang et al., 2026e). Instead of allowing free-form language outputs, this task requires a model to arrange shuffled candidate frames into their correct chronological order, reducing opportunities for linguistic shortcuts and forcing a genuine understanding of temporal progression. To tackle this challenging setting, we then introduce ChronoVision. Its central idea is to encourage the model to internally reconstruct the visual outcome of an event within its latent space, mirroring the human tendency to mentally imagine how a scene should unfold (Lee et al., 2025).

The training pipeline of ChronoVision consists of two main stages. During the supervised fine-tuning stage, we introduce the Reconstructive Visual Head. This module trains the model to use scattered information from the candidate frames to predict the latent representation of the final transformed state. Simultaneously, we incorporate a Region of Interest Attention Locating module. Guided by semantic locating cues, this module directs the model to precisely focus on the fine-grained dynamic key regions required for visual reasoning. Given long-horizon reasoning often leads to compounding errors (Motwani et al., 2025; Li et al., 2026m,b, 2025a, 2026c,b), we implement a reinforcement learning stage based on implicit process grounding. We design a Composite Reward Function that simultaneously evaluates outcome correctness, latent visual alignment, and attention focus. ChronoVision achieves a 74.8 percent in domain accuracy and a 71.6 percent out-of-domain accuracy on Vbvr-VQA, outperforming all compared open-source and proprietary models. Furthermore, ChronoVision achieves state-of-the-art performance on the challenging IntPhys2 benchmark (Bordes et al., 2025), with an overall accuracy of 55.0%. Overall, our contributions are summarized as follows:

- First, we introduce Vbvr-VQA. We redesign the video reasoning task as an image ordering task. Given an initial frame, a text prompt, and six shuffled frames, the model is required to accurately order the sequence. This forces the model to genuinely comprehend temporal evolution rather than relying on linguistic shortcuts.
- Second, we propose ChronoVision, a comprehensive framework for visual temporal reasoning. We introduce the Reconstructive Visual Head to predict the visual representation of the final transformed state within the internal latent space. We also introduce the Region of Interest Attention Locating module to guide the focus toward fine-grained dynamic key regions.
- Third, we incorporate a reinforcement learning stage based on implicit process grounding. We design a Composite Reward Function that comprehensively evaluates outcome correctness, latent visual alignment, and attention focus. This approach optimizes the internal reasoning trajectory and significantly improves robustness in long-horizon tasks.
- Fourth, extensive experiments validate the superiority of our approach. ChronoVision achieves state-of-the-art performance on Vbvr-VQA and attains the best overall accuracy on the challenging cross-domain IntPhys2 benchmark (Bordes et al., 2025).

2 Related Works

VLMs for Vision Reasoning. Recent VLMs shift from perception toward complex reasoning (Ke et al., 2025; Zhang et al., 2026e; Izadi et al., 2026), with continuous latent-space reasoning mitigating hallucination and language shortcuts (Qin et al., 2025; Zhang et al., 2025a; Li et al., 2026a, 2025d). GRPO-based reinforcement fine-tuning (Shao et al., 2024) with step-wise rewards and curricula (Zhang et al., 2025c; Deng et al., 2026; Xia et al., 2025; Yuan et al., 2025; Huang et al., 2026; Wang et al., 2026j) has driven progress on visual instruction tuning (Liu et al., 2023; Zhang et al., 2025d), yet a persistent gap remains between static spatial knowledge and temporal understanding (Feng et al., 2025; Cores et al., 2024; Shen et al., 2026b; Wang et al., 2026i; Shen et al., 2026c). We instead reformulate video reasoning as image ordering, grounding it in the latent visual representation of the correct final state.

Visual Cognition. VLMs show marked deficits in core knowledge, intuitive physics, and abstract reasoning relative to humans (Li et al., 2024; Schulze Buschoff et al., 2025; Shen et al., 2026a). Benchmarks such as MaRs-VQA (Cao et al., 2024), VRIQ (Khezresmaeilzadeh et al., 2026), Int-

Phys 2 (Bordes et al., 2025), and OmniSpatial (Jia et al., 2026) reveal perceptual and relational bottlenecks rather than pure language-logic limits, motivating latent imagery prediction as in Cognitive Supersensing (Li et al., 2026d). Matrix-reasoning benchmarks (RAVEN (John and Raven, 2003), I-RAVEN (Hu et al., 2021), CVR (Zerroug et al., 2022), RAVEN-IQ (Huang et al., 2023)) probe this gap spatially; we instead target it in the *temporal* domain by training the model to internally simulate the latent final visual state.

ROI Selection and Cropping. Tool-based approaches (Chain-of-Spot (Liu et al., 2024c), ZoomEye (Shen et al., 2025), Refocus (Fu et al., 2025)) and RL-optimized region exploration (DeepEyes (Zheng et al., 2026b), DeepEyesV2 (Hong et al., 2026)) improve evidence acquisition, while attention-driven methods (ICoT (Gao et al., 2025), ViCrop (Zhang et al., 2025b), FOCUS (Zhong et al., 2026)) avoid brittle coordinate prediction but remain sensitive to attention dispersion. We use semantically guided ROI attention locating to concentrate attention on dynamic key regions for temporal evolution modeling.

Reinforcement Learning. Recent RL advances improve stability and credit assignment (DAPO (Yu et al., 2026d), CPPO (Lin et al., 2026), VinePPO (Kazemnejad et al., 2024)) atop PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), with variants like Dr. GRPO (Liu et al., 2025d) correcting length bias. Unlike prior work optimizing textual or answer-level outcomes, we adopt GRPO for implicit process grounding via a composite reward over sequence correctness, latent visual alignment, and attention focus.

A comprehensive discussion of related work is provided in Appendix F.

3 Dataset and Benchmark

Current Visual Question Answering benchmarks primarily assess static image and text alignment alongside fundamental visual recognition. The community currently lacks datasets designed to evaluate abstract visual reasoning, including multistep transformation inference, causal temporal dynamics, and complex scene understanding. To address this limitation, we introduce Vbvr-VQA, built upon the open source Very Big Video Reasoning dataset (Wang et al., 2026e). These samples span diverse cognitive task categories and are partitioned into in domain and zero shot generalization splits. To explicitly measure the temporal logic and state tracking capabilities of the models, we reformulate the traditional multiple choice format into a strict cognitive visual sequence task.

This reformulation into an image-ordering task is a deliberate design choice rather than a limitation. Traditional visual question answering benchmarks frequently rely on multiple-choice formats or free-form text generation. These formats contain a critical flaw: models can exploit linguistic pattern-matching within the question stems or options to bypass actual visual comprehension. By enforcing a text-free, strict ordering formulation, we eliminate linguistic shortcuts and force the model to execute pure temporal and physical causal reasoning. Furthermore, while the output format is unified, the underlying physical scenarios are extremely diverse. The benchmark incorporates 100 distinct task generators covering fluid dynamics, kinematic collisions, and continuous spatial rotations. Testing models through a unified, strictly constrained interface across highly heterogeneous physical domains represents a standard paradigm for rigorous cognitive evaluation.

Under the standard evaluation pipeline, the model receives an initial video frame and a text prompt that serve as the visual premise and reasoning instruction. The model is then provided with six shuffled candidate images representing different evolutionary states of the visual transformation. The core objective is to predict the exact chronological permutation of these unordered frames. To construct this task, we uniformly divide the ground truth video along the temporal axis into six intervals of equal length. We extract the final frame of each segment to form a temporal sequence that depicts the progressive transformation. We subsequently apply a random permutation to these frames to generate the unordered candidate set.

To support fine grained visual reasoning and mitigate the common grounding perception mismatch problem in Vision Language Models, we augment Vbvr-VQA with dense annotations. Based on the ground truth sequence, the annotation pipeline synthesizes intermediate reasoning traces to explicitly identify the dynamic transformation regions. This process generates semantic visual cues formatted as `<LOCATE>...</LOCATE>`. Simultaneously, the pipeline provides precise bounding box coordinates for the regions undergoing state changes. These coordinates serve as dense spatial supervision for the downstream attention modules. Figure 1 presents a

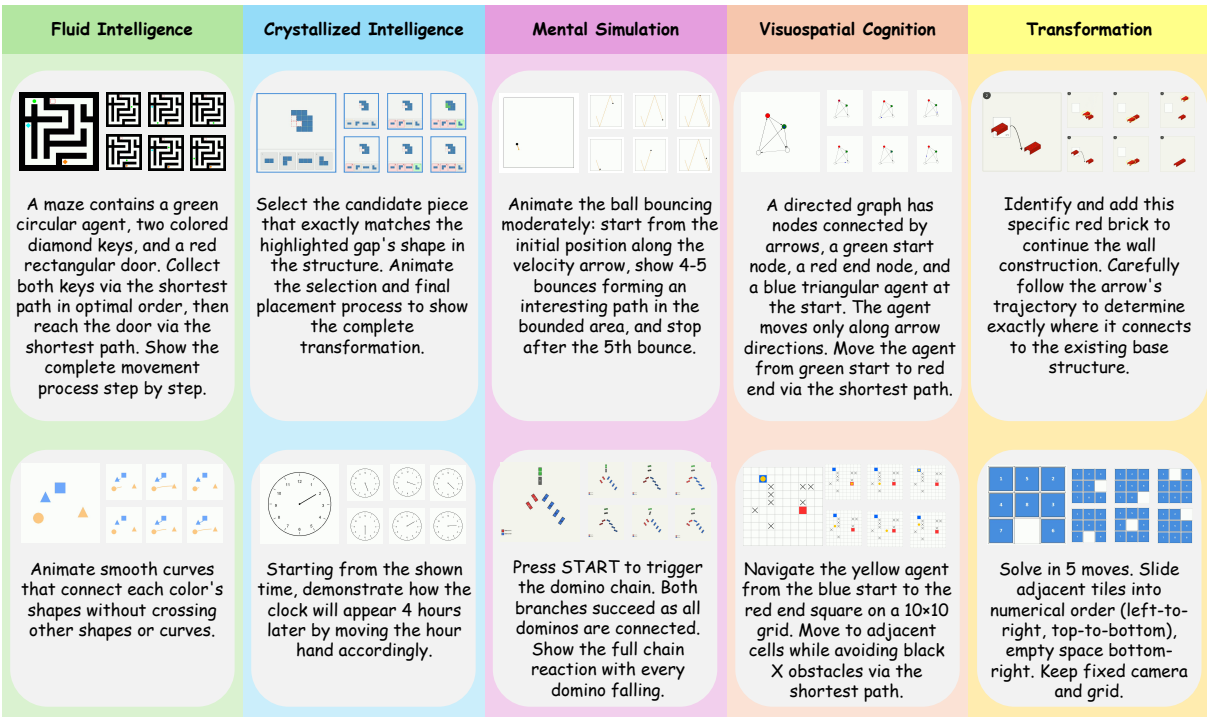

Figure 1: Examples of the Vbvr-VQA dataset

set of representative examples from the Vbvr-VQA benchmark to illustrate the diversity of the visual reasoning tasks. By formulating the evaluation as a sequence ordering task rather than a single choice selection, Vbvr-VQA strictly requires the model to understand physical trajectories and causal relationships. This design effectively prevents the models from exploiting shortcuts based on superficial single frame matching. We provide additional information in Appendix A.

4 Methodology

4.1 Preliminaries and Problem Formulation

We aim to train a multimodal large language model, parameterized by θ, to model the conditional distribution $\pi_\theta(y \mid x)$ for a multimodal input $x = (\mathcal{V}, \mathcal{Q})$. Here, y is the predicted chronological permutation. The visual input $\mathcal{V}$ includes an initial query image v_{query} and a candidate set $\mathcal{O} = \{o_1, \ldots, o_6\}$ sampled from distinct temporal points. The textual prompt $\mathcal{Q}$ specifies the reasoning task. A visual encoder $\text{Encoder}_{\text{vis}}(\cdot)$ extracts dense features from $\mathcal{V}$, which are projected into visual tokens. The language model backbone $\text{Encoder}_{\text{txt}}(\cdot)$ processes the concatenated visual and text tokens to generate y autoregressively.

Figure 2 provides an overview of ChronoVision. We next describe its three main components: the Reconstructive Visual Head, the ROI Attention Locating module, and the reinforcement learning stage with implicit process grounding.

4.2 Supervised Fine-Tuning with Reconstructive Visual Head

To acquire advanced visual cognitive capabilities, we introduce the Reconstructive Visual Head (RVH). This auxiliary module predicts the latent representation of the final transformed state, which corresponds to the final chronological image, from the shuffled candidate frames.

Let $v_{\text{final}} \in \mathcal{O}$ denote the candidate image corresponding to the final chronological state. We use its unpooled sequence embedding from the frozen visual encoder as the RVH supervision target:

$$h_{\text{final}} = \text{Encoder}_{\text{vis}}(v_{\text{final}}) \in \mathbb{R}^{N_{\text{img}} \times d},$$

where N_{img} is the number of visual tokens for each image.

The RVH module $g_\psi(\cdot)$ is a two-layer multilayer perceptron operating in parallel with the text decoder. Let $\mathbf{H}_{\text{opt}} \in \mathbb{R}^{M \times d}$ denote the hidden states of the visual tokens corresponding to the shuffled candidate images produced by $\text{Encoder}_{\text{txt}}(\cdot)$. The RVH predicts the final-state latent representation as

$$\hat{h}_{\text{final}} = g_\psi(\mathbf{H}_{\text{opt}}), \qquad \hat{h}_{\text{final}} \in \mathbb{R}^{N_{\text{img}} \times d}.$$

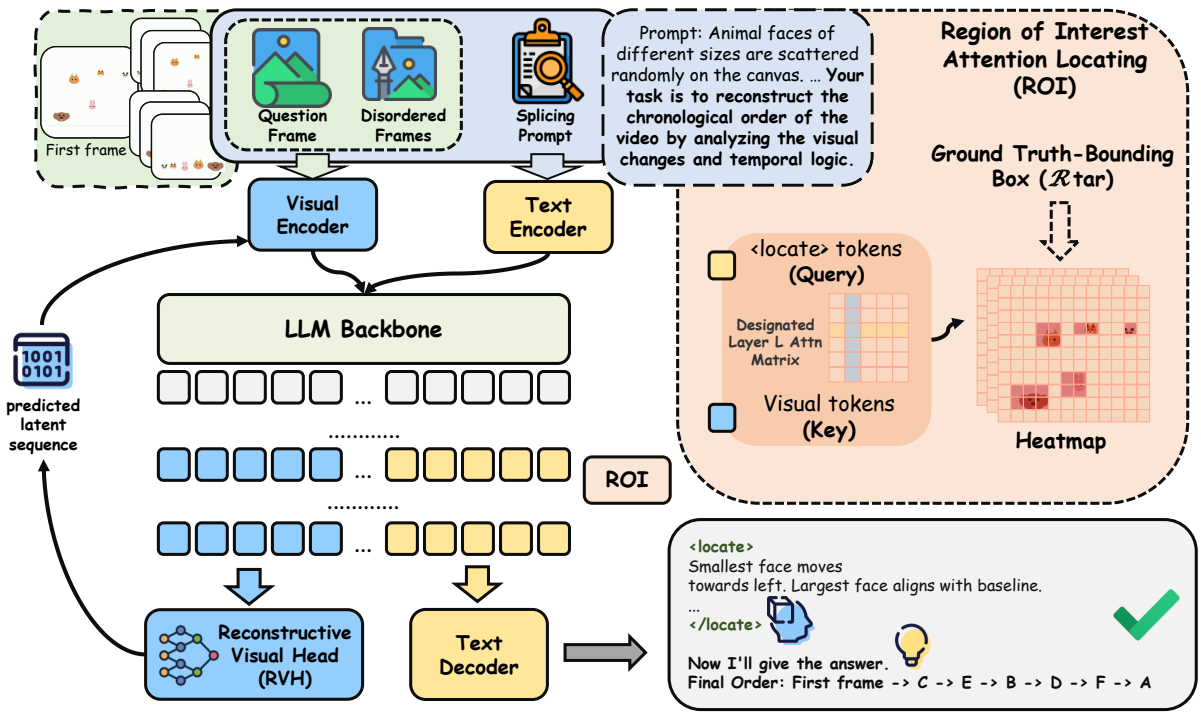

Figure 2: Overall pipeline of ChronoVision. Given a query frame, shuffled candidate frames, and a task prompt, the model encodes visual and textual inputs with a unified backbone. On top of this backbone, the Reconstructive Visual Head predicts the latent representation of the final transformed state, while the ROI Attention Locating module guides attention toward dynamic evidence regions for temporal reasoning.

We train the RVH using a Mean Squared Error (MSE) loss together with the standard autoregressive cross-entropy loss for the target sequence y^*:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|y^*|} \log \pi_\theta(y_t^* \mid x, y_{<t}^*) + \lambda_{\text{RVH}} \cdot \text{MSE}\left(\hat{h}_{\text{final}}, h_{\text{final}}\right). \tag{1}$$

Here, λ_{RVH} balances the autoregressive and latent reconstruction objectives. This supervision connects sequence prediction with the visual representation of the final state.

4.3 Supervised Fine-Tuning with Region of Interest Attention Locating

Resolving abstract visual transformations requires tracking localized dynamic changes. To prevent scattered attention and semantic noise, we propose a Region of Interest (ROI) Attention Locating module. Guided by dataset annotations, we prompt the model to generate a concise semantic locate cue enclosed by specific structural tokens before outputting the sequence. This isolates the key visual evidence.

To prevent spatial signal fragmentation, we condense the attention into a designated intermediate layer ℓ. We extract the text-to-image self-attention weights at layer ℓ using the generated locate tokens as queries and the initial query image tokens as keys. Using the ground-truth bounding box $\mathcal{R}_{\text{tar}}$ of the transformation area, we optimize spatial alignment via an Attention Condensation loss:

$$\mathcal{L}_{AC} = -\log\left(s^{(\ell)}(\mathcal{R}_{\text{tar}})\right) \tag{2}$$

where $s^{(\ell)}(\mathcal{R}_{\text{tar}})$ is the normalized mean-attention ratio within the bounding box:

$$s^{(\ell)}(\mathcal{R}_{\text{tar}}) = \frac{\frac{1}{|\mathcal{R}_{\text{tar}}|}\sum_{p\in\mathcal{R}_{\text{tar}}} \tilde{a}_p^{(\ell)}}{\frac{1}{|\mathcal{R}_{\text{img}}|}\sum_{p\in\mathcal{R}_{\text{img}}} \tilde{a}_p^{(\ell)}} \tag{3}$$

Here, $\mathcal{R}_{\text{img}}$ is the set of all image tokens, and $\tilde{a}_p^{(\ell)}$ is the aggregated attention weight at position p in layer ℓ, averaged over all query tokens and attention heads. The overall objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SFT}} + \alpha\mathcal{L}_{AC}$.

A detailed methodological discussion on why we prioritize attention alignment over explicit bounding box coordinate regression is provided in B.2.

4.4 Reinforcement Learning with Implicit Process Grounding

To mitigate compounding errors in long-horizon reasoning, we introduce a reinforcement learning stage based on Group Relative Policy Optimization (GRPO) (Shao et al., 2024). We design an implicit process grounding mechanism that decouples the reward function into three granular components bounded within $[0, 1]$, evaluating the reasoning trajectory without human annotation.

1. Final Outcome Reward (R_{out}). We assign a sparse reward based on the exact match between the predicted sequence y and the ground-truth sequence y^*:

$$R_{\text{out}} = \begin{cases} 1, & \text{if } y = y^* \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

2. Latent-Grounding Process Reward (R_{latent}). We evaluate the reasoning process at the sentence level to provide dense supervision. Let K be the total number of reasoning steps (sentences) in y. At step k, we feed the contextualized hidden states of the visual tokens into the RVH to generate a latent spatial feature sequence $Z_k \in \mathbb{R}^{N_{\text{img}} \times d}$. Simultaneously, the visual encoder provides candidate visual features $\mathbf{V} = \{V_i\}_{i=1}^{6}$, where $V_i = \text{Encoder}_{\text{vis}}(o_i) \in \mathbb{R}^{N_{\text{img}} \times d}$ corresponds to the shuffled options. The process reward measures the maximum cosine similarity between the reasoning intent Z_k and the available visual features V_i:

$$R_{\text{latent}} = \frac{1}{K} \sum_{k=1}^{K} \max_{V_i \in \mathbf{V}} \left(\frac{\cos(Z_k, V_i) + 1}{2} \right), \tag{5}$$

where $\cos(\cdot, \cdot)$ denotes the flattened cosine similarity. This anchors textual reasoning to underlying image features.

3. Unsupervised Visual Focus Reward (R_{focus}). To prevent attention collapse, we introduce an unsupervised focus reward based on internal self-attention maps. Let $\mathbf{a}_k \in \mathbb{R}^{N_{\text{img}}}$ be the normalized self-attention probability distribution over the N_{img} visual tokens at step k. Using its Shannon entropy $H(\mathbf{a}_k) = -\sum_{i=1}^{N_{\text{img}}} a_{k,i} \log a_{k,i}$, we formulate a normalized negative-entropy reward:

$$R_{\text{focus}} = \frac{1}{K} \sum_{k=1}^{K} \left(1 - \frac{H(\mathbf{a}_k)}{\log N_{\text{img}}} \right). \tag{6}$$

GRPO Optimization Objective. The total reward is $R_{\text{total}} = \omega_1 R_{\text{out}} + \omega_2 R_{\text{latent}} + \omega_3 R_{\text{focus}}$, where ω_1, ω_2, and ω_3 are balancing weights.

For each prompt $x \sim \mathcal{D}$, the old policy π_{old} samples G outputs $\{y_1, \ldots, y_G\}$. We compute the group-relative advantage $\hat{A}_i = (R_{\text{total}}^{(i)} - \mu_R)/\sigma_R$, where μ_R and σ_R are the mean and standard deviation of the group rewards. The policy π_θ is optimized using a clipped surrogate objective with a Kullback-Leibler penalty:

$$\begin{aligned} \mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D},\, y \sim \pi_{\text{old}}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Bigg(& \\ \min \Big(r_i \hat{A}_i,\ \text{clip}(r_i,\ 1-\epsilon,\ 1+\epsilon) \hat{A}_i \Big) & \\ - \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)} \Bigg) \Bigg], & \end{aligned} \tag{7}$$

where $r_i = \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)}$.

5 Experiments

5.1 Experimental Setup

Implementation Details. We evaluate ChronoVision on the Vbvr-VQA benchmark to test fine-grained visual understanding and logical reasoning capabilities in visual temporal transformations. More implementation details will be in Appendix B. We report the Exact Match accuracy as the primary evaluation metric. A prediction is considered correct if and only if the entire generated sequence of the six candidate frames exactly matches the ground-truth chronological order.

Baselines. We benchmark ChronoVision against state-of-the-art open-source and proprietary vision-language models. The open-source baselines, selected for their strong spatial reasoning and recognition capabilities, include Qwen 3.5 397B (Qwen Team, 2026), GLM-4.6V (Hong et al., 2025), and Qwen 3.5 9B (Qwen Team, 2026). To establish a reference for standard zero-shot capabilities, we also evaluate widely used commercial models, specifically Gemini 3.1 Pro (Google Deepmind, 2026), Gemini 3.0 Flash (Google Deepmind, 2025), Claude Opus 4.6 (Anthropic, 2026), GPT o3 (OpenAI, 2025), and GPT-5.4 (OpenAI, 2026).

5.2 Main Results

Performance on Vbvr-VQA. As shown in Table 1, ChronoVision establishes a new state-of-the-art on the Vbvr-VQA benchmark, achieving 74.8% accuracy on the In-Domain split, 71.6% on the Out-of-Domain split, and 73.2% overall. It surpasses the strongest proprietary baseline, Claude

Table 1: Benchmarking results on Vbvr-VQA benchmark. Overall In-Domain (ID) and Out-of-Domain (OOD) exact-match accuracy is reported alongside category-wise performance. Higher is better. **Bold**: best in group; underline: second best.

		In-Domain by Category						Out-of-Domain by Category					
Models	**Overall**	**Avg.**	**Flu.**	**Cry.**	**Vis.**	**Men.**	**Trans.**	**Avg.**	**Flu.**	**Cry.**	**Vis.**	**Men.**	**Trans.**
Qwen 3.5 9B (Qwen Team, 2026)	14.2	16.8	10.8	13.3	20.0	40.0	12.7	11.6	13.3	20.0	5.0	12.0	10.0
GLM-4.6V (Hong et al., 2025)	21.4	19.2	10.8	22.2	21.7	20.0	23.6	23.6	15.6	30.0	25.0	16.0	26.0
GPT-5.4 (OpenAI, 2026)	33.8	28.0	23.1	31.1	25.0	32.0	32.7	39.6	20.0	47.5	30.0	40.0	49.0
Gemini 3.1 Pro (Google Deepmind, 2026)	41.8	52.8	32.3	66.7	66.7	60.0	47.3	30.8	42.2	62.5	47.5	48.0	2.0
GPT o3 (OpenAI, 2025)	46.0	42.8	26.2	60.0	45.0	64.0	36.4	49.2	31.1	62.5	45.0	60.0	51.0
Gemini 3.0 Flash (Google Deepmind, 2025)	46.6	47.2	30.8	51.1	53.3	**80.0**	41.8	46.0	31.1	37.5	45.0	48.0	56.0
Qwen 3.5 397B (Qwen Team, 2026)	49.4	45.6	21.5	60.0	58.3	44.0	49.1	53.2	42.2	67.5	47.5	44.0	57.0
Claude Opus 4.6 (Anthropic, 2026)	55.8	50.8	30.8	66.7	60.0	56.0	49.1	60.8	48.9	67.5	42.5	56.0	72.0
ChronoVision	**73.2**	**74.8**	**72.3**	**77.8**	**75.0**	68.0	**78.2**	**71.6**	**66.7**	**75.0**	**72.5**	**64.0**	**74.0**

Opus 4.6 (50.8% ID, 60.8% OOD). Our 9B parameter framework successfully outperforms models with significantly larger parameter scales, including Qwen 3.5 397B and GPT o3. The Figure 3 shows the visualization of qualitative comparison of reasoning chains on a transformation planning problem from Vbvr-VQA. And more examples will be in Appendix E

General Vision-Language Capabilities. We evaluate ChronoVision on 7 standard multimodal benchmarks. As presented in Table 2, ChronoVision maintains performance strictly comparable to the vanilla Qwen 3.5 9B across all general tasks (e.g., 78.8 vs. 78.4 on MMMU, 85.9 vs. 85.7 on MathVista), confirming that our attention locating and reconstructive objectives enhance spatial-temporal reasoning without degrading general visual-language understanding.

5.3 Ablation Studies

We conduct ablation studies to isolate the contributions of our architectural components and RL reward designs, as detailed in Table 3. We provide additional, detailed ablation studies in Appendix C covering partial match evaluation, the evolution of latent representations, the impact of chain-of-thought reasoning, prefix sensitivity, intermediate state perturbations, latent sequence interventions, linear probing, and comparisons with related temporal reasoning methods.

Effectiveness of Training Processes. Starting from the baseline model without the Reconstructive Visual Head (*SFT w/o head*), adding the RVH (*SFT only*) yields a direct +3.2% ID and +2.8% OOD improvement. This confirms that explicitly predicting latent visual states effectively grounds sequence ordering. Integrating the ROI Attention Locating module (*SFT with ROI*) further improves overall accuracy by +1.2% (from 69.0% to 70.2%), proving its ability to filter spatial semantic noise. Finally, applying the GRPO reinforcement learning stage (*SFT with ROI + RL*) provides another substantial boost (+3.2% ID, +2.8% OOD), demonstrating that RL effectively suppresses error accumulation during long-horizon temporal reasoning.

Ablation on Reward Components. Within the RL stage, we ablate the composite reward function to verify that all components are indispensable. Removing the sparse *Answer Reward* (outcome reward R_{out}) causes the most severe performance drop (-3.4% Overall). Crucially, removing the dense intermediate signals also degrades performance: omitting the *Latent-Grounding Process Reward* (latent alignment process reward R_{latent}) leads to a -2.2% overall decline, while removing the *ROI Reward* (unsupervised visual focus reward R_{focus}) results in a -1.4% overall drop. This verifies that both latent trajectory guidance and forced spatial attention are critical for maximizing reasoning accuracy.

5.4 Out-of-Domain Evaluation on Physical Dynamics

While our in-domain evaluation focused on abstract cognitive reasoning (e.g., spatial puzzles and block transformations), we further evaluate our model's generalization capabilities on real-world physical dynamics using the highly challenging IntPhys 2 benchmark (Bordes et al., 2025). Rather than abstract shapes, this benchmark requires models to sequence frames depicting realistic physical events based on causal logic and physical plausibility (e.g., gravity, collisions, and momentum). Examples of these realistic scenarios are illustrated in Figure 4.

Evaluating intuitive physics is notoriously difficult for current Vision-Language Models (VLMs),

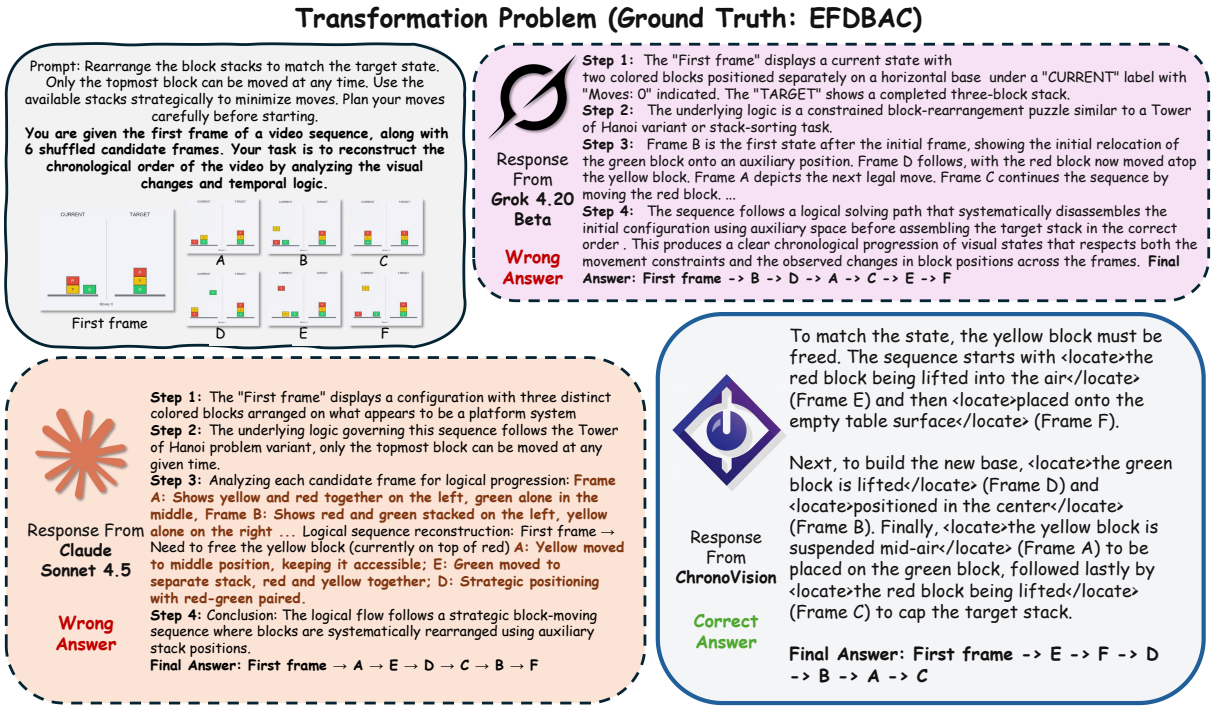

Figure 3: Qualitative comparison of reasoning chains on a transformation planning problem from Vbvr-VQA. The task requires reconstructing the correct chronological order of six shuffled candidate frames by inferring valid intermediate moves under the top-block-only constraint.

Table 2: **General Vision-Language Understanding Results.** Best performance in **bold**.

Models	MMMU	MMBench	MMStar	MathVista	AI2D	HallusionBench	BabyVision
Qwen3.5 9B (Qwen Team, 2026)	78.4	**90.1**	**79.7**	85.7	**90.2**	69.3	25.8
ChronoVision	**78.8**	88.2	78.5	**85.9**	**90.2**	**69.6**	**27.3**

with most baselines hovering near random chance (approximately 50%). However, as detailed in Table 4, ChronoVision achieves consistent improvements across the Easy, Medium, and Hard subsets of IntPhys 2.

Overall, our model attains an accuracy of 55.0%, yielding a +6.5% absolute gain over the strong base model, Qwen 3.5 9B (48.5%). The improvement is particularly pronounced on the Easy subset, where ChronoVision reaches 62.5%, outperforming the baseline by +11.5%. These substantial OOD gains confirm that our framework enables the model to internalize generalized physical laws and spatial-temporal continuity.

We also evaluate ChronoVision on two real-world video reasoning benchmarks, Video-Holmes and LongVideo-Reason, where it obtains overall scores of 45.89 and 74.7, respectively. The complete results are reported in Appendix D.

6 Related Works

Our work is closely related to VLMs for Vision Reasoning, Visual Cognition, ROI Selection and Cropping and Reinforcement learning. A comprehensive discussion is provided in Appendix F

7 Conclusion

This paper introduces ChronoVision, a framework designed to resolve the text bottleneck in long horizon abstract visual reasoning. By integrating a Reconstructive Visual Head and a Region of Interest Attention module during supervised fine tuning, the model acquires implicit visual reasoning capabilities and fine grained spatial focus. We further propose a reinforcement learning stage utilizing a composite reward function to mitigate compounding errors in textual chains of thought. Evaluated on our proposed Vbvr-VQA benchmark and the cross domain IntPhys2 dataset, ChronoVision achieves state of the art performance against open source and proprietary models. Future research will ex-

Table 3: Ablation study results on training processes and rewards (Accuracy %). Overall In-Domain and Out-of-Domain accuracy metrics are reported alongside category-wise performance. Higher is better. **Bold**: best in group; underline: second best.

		In-Domain by Category						Out-of-Domain by Category					
Models	**Overall**	**Avg.**	**Flu.**	**Cry.**	**Vis.**	**Men.**	**Trans.**	**Avg.**	**Flu.**	**Cry.**	**Vis.**	**Men.**	**Trans.**
Training Processes													
SFT w/o head	66.0	66.8	66.2	73.3	66.7	56.0	67.3	65.2	62.2	70.0	65.0	52.0	68.0
SFT only	69.0	70.0	69.2	75.6	68.3	60.0	72.7	68.0	64.4	72.5	67.5	56.0	71.0
SFT with ROI	70.2	71.6	69.2	75.6	71.7	64.0	74.5	68.8	64.4	72.5	70.0	56.0	72.0
SFT with ROI + RL	**73.2**	**74.8**	**72.3**	**77.8**	**75.0**	**68.0**	**78.2**	**71.6**	**66.7**	**75.0**	**72.5**	**64.0**	**74.0**
Reward Components													
w/o Answer Reward	69.8	71.2	69.2	75.6	70.0	64.0	74.5	68.4	64.4	72.5	70.0	56.0	71.0
w/o Latent-Grounding Reward	71.0	72.8	70.8	75.6	71.7	**68.0**	76.4	69.2	64.4	72.5	70.0	60.0	72.0
w/o ROI Reward	71.8	74.0	70.8	**77.8**	73.3	**68.0**	**78.2**	69.6	64.4	72.5	70.0	60.0	73.0
Full Reward	**73.2**	**74.8**	**72.3**	**77.8**	**75.0**	**68.0**	**78.2**	**71.6**	**66.7**	**75.0**	**72.5**	**64.0**	**74.0**

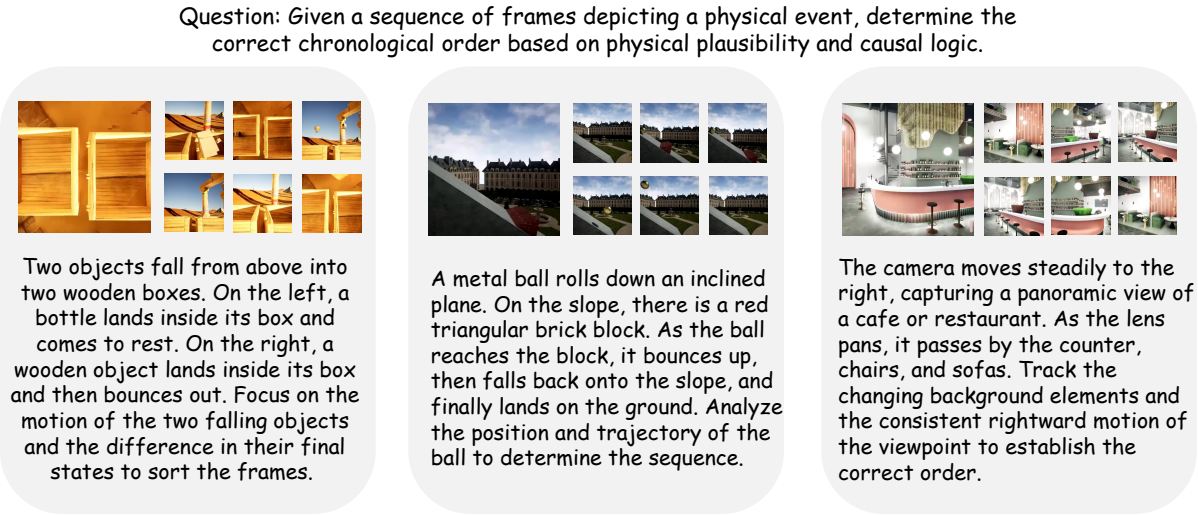

Figure 4: Examples of out-of-domain physical reasoning. We visualize three distinct real-world scenarios where the model must reconstruct the chronological order based on physical plausibility and causal logic. **Left:** Two objects fall into boxes exhibiting different kinetic behaviors (one resting, one bouncing), requiring the model to infer collision outcomes. **Middle:** A metal ball rolls down a slope, hits a stationary block, and bounces, testing intuitive understanding of momentum and trajectory. **Right:** A camera pans steadily across a café, challenging the model to track spatial consistency and continuous viewpoint shifts in a complex real-world environment.

Table 4: **Generalization on IntPhys 2.** Out-of-domain performance evaluating cognitive understanding of universal physical laws. Accuracy (%) is reported. Best performance in **bold**.

Models	Easy	Medium	Hard	Overall
Qwen 3.5 9B (Qwen Team, 2026)	51.0	48.5	47.6	48.5
ChronoVision	**62.5**	**54.8**	**53.0**	**55.0**

plore integrating generative models to explicitly visualize internal visual chains of thought.

Limitation

While ChronoVision demonstrates strong temporal reasoning and visual cognition capabilities on Vbvr-VQA and IntPhys2, the current study primarily focuses on validating the effectiveness of latent sequence reconstruction and process grounding within a moderate-scale 9B MLLM setting. Future work could further investigate scalability to larger multimodal backbones and larger and more diverse training corpora. Besides, the current framework relies on dense auxiliary supervision, including semantic locating cues and spatial bounding-box annotations, to guide the ROI Attention Locating module. Future research may explore weaker-supervision or annotation-free alternatives to improve scalability and adaptability across broader video reasoning domains.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Anthropic. 2026. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. 2025. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*.

Fanpu Cao, Xin Zou, Xuming Hu, and Hui Xiong. 2026. When looking is not enough: Visual attention structure reveals hallucination in mllms. *arXiv preprint arXiv:2605.11559*.

Xu Cao, Yifan Shen, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Jintai Chen, Meihuan Huang, Jianguo Cao, Aidong Zhang, and 1 others. 2024. What is the visual cognition gap between humans and multimodal llms? *arXiv preprint arXiv:2406.10424*.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Lekai Chen, Ashutosh Trivedi, and Alvaro Velasquez. 2024. Llms as probabilistic minimally adequate teachers for dfa learning. *arXiv preprint arXiv:2408.02999*.

Lekai Chen, Ashutosh Trivedi, and Alvaro Velasquez. 2026a. Towards persistent noise-tolerant active learning of regular languages with class query. In *The Fourteenth International Conference on Learning Representations*.

Lekai Chen, Alvaro Velasquez, and Ashutosh Trivedi. 2025b. Cedar: A counter-example driven agent with regular restriction.

Qianyu Chen, Ziteng Feng, Canran Xiao, and Runxuan Tang. 2026b. Inducekv: Fixed-footprint continual adaptation of multimodal llms via inducing kv memories. *arXiv preprint arXiv:2607.02010*.

Qianyu Chen, Canran Xiao, and Runxuan Tang. 2026c. Hidden forgetting in continual multimodal learning: When accuracy survives but grounding fails. *arXiv preprint arXiv:2607.02020*.

Yihang Chen, Pin Qian, Su Wang, Sipeng Zhang, Huan Xu, Shuhuai Lin, and Xinpeng Wei. 2026d. Does rag know when retrieval is wrong? diagnosing context compliance under knowledge conflict. *arXiv preprint arXiv:2605.14473*.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2026e. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. 2024. Lost in time: A new temporal benchmark for videollms. *arXiv preprint arXiv:2410.07752*.

Chengxiao Dai, Zhanhui Lin, Zhaokun Yan, Youyang Ni, Chenjun Lei, and Luyan Zhang. 2026a. Coordinating from memory: Graph-structured experience reuse for multi-agent adaptation in dynamic manufacturing. *arXiv preprint arXiv:2607.19985*.

Chengxiao Dai, Yue Xiu, and Dusit Niyato. 2025. From graph retrieval to schema realization: Counterfactual validation for text-to-sparql over heterogeneous knowledge graphs. *arXiv preprint arXiv:2508.01815*.

Chengxiao Dai, Zhaokun Yan, Chenjun Lei, Qiao Li, and Luyan Zhang. 2026b. Safe remediation as risk-constrained intervention decision in microservice systems. *arXiv preprint arXiv:2607.20005*.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2026. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *Advances in Neural Information Processing Systems*, 38:123817–123846.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. 2025. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*.

François Fleuret, Ting Li, Charles Dubout, Emma K Wampler, Steven Yantis, and Donald Geman. 2011. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. 2025. Refocus: Visual editing as a chain of thought for structured image understanding. *arXiv preprint arXiv:2501.05452*.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19520–19529. IEEE.

Google Deepmind. 2025. Gemini 3 flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Google Deepmind. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Hanghui Guo, Jia Zhu, Changfan Pan, Qing Wang, Cong Zhou, and Chaojun Meng. 2026. Ke-lpg: Toward semantic refinement for lesson plan generation. *Arabian Journal for Science and Engineering*, 51(5):6761–6774.

Xinjie He, Zhiyuan Lin, Su Liu, Jialun Wu, Qiyang Xie, Weikai Zhou, and Shuai Xiao. 2026. What training data teaches rl memory agents: An empirical study of curriculum effects in memory-augmented qa. *arXiv preprint arXiv:2605.23067*.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, and Guohai Xu. 2026. Deepeyesv2: Toward agentic multimodal model. In *International Conference on Learning Representations*, volume 2026, pages 114851–114872.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Sheng Hu, Yuqing Ma, Xianglong Liu, Yanlu Wei, and Shihao Bai. 2021. Stratified rule-aware network for abstract visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1567–1574.

Jiaxing Huang, Jingyi Zhang, Kai Jiang, Han Qiu, Xiaoqin Zhang, Ling Shao, Shijian Lu, and Dacheng Tao. 2025. Visual instruction tuning towards general-purpose multimodal large language model: A survey. *International Journal of Computer Vision*, 133(11):8151–8189.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.

Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2026. Spotlight on token perception for multimodal reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 103291–103325.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Vahedi, Hosein Hasani, and Mahdieh Baghshah. 2026. Visual structures help visual reasoning: Addressing the binding problem in lvlms. *Advances in Neural Information Processing Systems*, 38:162930–162968.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. 2026. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, volume 2026, pages 35634–35670.

Xueqian Jiang, Jialun Wu, and Jiabei Liu. 2026. A tiered spatial validation workflow for geographic prediction models. *MethodsX*, page 104072.

Qirui Jin, Lingching Tung, Kenan Li, Qiyang Shi, Yushi She, Huanzhong Jia, Harrison Zhao, Kejing Xia, Zhenbang Du, Yikai Zhang, and 1 others. 2026a. Chainswe: Benchmarking coding agents on multi-bug software maintenance. *arXiv preprint arXiv:2607.02606*.

Zichun Jin, Zihan Zhou, Yinan Liu, Bin Wang, and Xiaochun Yang. 2026b. Adapting knowledge graphs for behavior denoising in sequential recommendation. *Preprint*, arXiv:2608.21243.

John and Jean Raven. 2003. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer.

Zhehan Kan, Xinghua Jiang, Yanlin Liu, Xiaochen Yang, Zhixiang Wei, Shifeng Liu, Yubo Zhu, Qingmin Liao, Wenming Yang, Xin Li, and 1 others. 2026. Uvu: Improving multimodal understanding via vision-language unified autoregressive paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26230–26239.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. 2024. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, and 1 others. 2025. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*.

Tina Khezresmaeilzadeh, Jike Zhong, and Konstantinos Psounis. 2026. Vriq: Benchmarking and analyzing visual-reasoning iq of vlms. *arXiv preprint arXiv:2602.05382*.

Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. 2025. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9241–9251.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. 2026a. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pages 148076–148090.

Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Hao Sun, and Junyu Gao. 2026b. Exploring efficient open-vocabulary segmentation in the remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5982–5991.

Bingyu Li, Feiyu Wang, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. 2026c. Maris: Marine open-vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24939–24949.

Bingyu Li, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. 2025a. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1308–1317.

Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, and 1 others. 2026d. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*.

Fan Li, Mengting Pan, Sijia Xu, Xiaoyang Wang, Chen Chen, and Wenjie Zhang. 2026e. Answer-then-edit: Reasoning skeleton editing for anti-distillation with preserved utility. *arXiv preprint arXiv:2607.20440*.

Jialin Li, Zhenhao Chen, Hanjun Luo, and Hanan Salam. 2026f. Prefix: Understand and adapt to user preference in human-agent interaction. *arXiv preprint arXiv:2602.06714*.

Kenan Li, Qirui Jin, Liao Zhu, Xiaosong Huang, Yijia Wu, Yikai Zhang, Xin Zhang, Zijian Jin, Yufan Huang, Elsie Nallipogu, and 1 others. 2026g. Oracle-swe: Quantifying the contribution of oracle information signals on swe agents. *arXiv preprint arXiv:2604.07789*.

Kenan Li, Rongzhi Li, Linghao Zhang, Qirui Jin, Liao Zhu, Xiaosong Huang, Geng Zhang, Yikai Zhang, Shilin He, Chengxing Xie, and 1 others. 2026h. Repolaunch: Automating build&test pipeline of code repositories on any language and any platform. *arXiv e-prints*, pages arXiv–2603.

Ningyuan Li, Junrui Liu, Yi Shan, Minghui Huang, Ziren Gong, and Tong Li. 2026i. Pankrag: Enhancing graph retrieval via globally aware query resolution and dependency-aware reranking mechanism. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19157–19161. IEEE.

Songze Li, Xiaoke Guo, Tianqi Liu, Biao Yi, Zhaoyan Gong, Zhiqiang Liu, Huajun Chen, and Wen Zhang. 2026j. What's missing in screen-to-action? towards a ui-in-the-loop paradigm for multimodal gui reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 17674–17690.

Songze Li, Yarong Lan, Zhongpu Bo, Zhaoyang Wang, Zhiqiang Liu, Yuan Yuan, Chengtao Gan, Menghao Qian, Enpei Niu, Xiaoke Guo, and 1 others. 2026k. Scaling llm knowledge boundaries via distribution-optimized synthesis. *arXiv preprint arXiv:2606.23271*.

Songze Li, Zhiqiang Liu, Zhaoyan Gong, Xiaoke Guo, Zhongpu Bo, Zhengke Gui, Lei Liang, Huajun Chen, and Wen Zhang. 2025b. Last layer logits to logic: Empowering llms with logic-consistent structured knowledge reasoning. *arXiv preprint arXiv:2511.07910*.

Songze Li, Zhiqiang Liu, Zhengke Gui, Huajun Chen, and Wen Zhang. 2025c. Enrich-on-graph: Query-graph alignment for complex reasoning with llm enriching. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7683–7703.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. 2026l. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7460.

Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, and 1 others. 2024. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*.

Zhenyu Li, Guanlin Wu, Cheems Wang, and Yongqiang Zhao. 2026m. Limited reasoning space: The cage of long-horizon reasoning in llms. *arXiv preprint arXiv:2602.19281*.

Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi Li, Jian Yang, Chenghua Lin, and Min Zhang. 2026n. Dynamics within latent chain-of-thought: An empirical study of causal structure. *arXiv preprint arXiv:2602.08783*.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and 1 others. 2025d. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*.

Jia Liang and Liangming Pan. 2026. Do latent-cot models think step-by-step? a mechanistic study on sequential reasoning tasks. *arXiv preprint arXiv:2602.00449*.

Ruqi Liao, Chuqing Zhao, Jin Li, Weiqi Feng, Yi Lyu, Bingxian Chen, and Haochen Yang. 2025. Catp: Cross-attention token pruning for accuracy preserved multimodal model inference. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, pages 1100–1104. IEEE.

Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. 2026. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *Advances in Neural Information Processing Systems*, 38:61043–61068.

Bingchen Liu, Yuanyuan Fang, Lei Liu, Guangyuan Dong, Xing Fu, Yuanyuan Gao, Shuyue Wei, Xin Li, and Xiangtian Meng. 2026a. Conmem: Contribution-aware memory for long-horizon manufacturing inspection logs. *arXiv preprint arXiv:2607.28126*.

Hanzhi Liu, Ningkang Peng, Qianyu Chen, Canran Xiao, and 1 others. 2026b. Affordance-first decomposition for continual learning in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3908–3919.

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

Jiachen Liu, Jiaxin Pei, Jintao Huang, Chenglei Si, Ao Qu, Xiangru Tang, Runyu Lu, Lichang Chen, Xiaoyan Bai, Haizhong Zheng, and 1 others. 2026c. The last human-written paper: Agent-native research artifacts. *arXiv preprint arXiv:2604.24658*.

Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. 2026d. Dual-pathway circuits of object hallucination in vision-language models. *arXiv preprint arXiv:2605.13156*.

Jun Liu, Zhenglun Kong, Peiyan Dong, Xuan Shen, Pu Zhao, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Xue Lin, and 1 others. 2025a. Rora: Efficient fine-tuning of llm with reliability optimization for rank adaptation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jun Liu, Zhenglun Kong, Peiyan Dong, Changdi Yang, Tianqin Li, Yanyue Xie, Yifan Gong, Xuan Shen, Pu Zhao, Hao Tang, and 1 others. 2026e. Structured agent distillation for large language model agents. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pages 3676–3685.

Jun Liu, Zhenglun Kong, Changdi Yang, Fan Yang, Tianqi Li, Peiyan Dong, Joannah Nanjekye, Hao Tang, Geng Yuan, Wei Niu, and 1 others. 2025b. Rcr-router: Efficient role-aware context routing for multi-agent llm systems with structured memory. *arXiv preprint arXiv:2508.04903*.

Jun Liu, Zhenglun Kong, Pu Zhao, Changdi Yang, Xuan Shen, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Xue Lin, and 1 others. 2025c. Toward adaptive large language models structured pruning via hybrid-grained weight importance assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18879–18887.

Jun Liu, Zhenglun Kong, Pu Zhao, Weihao Zeng, Hao Tang, Xuan Shen, Changdi Yang, Wenbin Zhang, Geng Yuan, Wei Niu, and 1 others. 2024a. Tsla: A task-specific learning adaptation for semantic segmentation on autonomous vehicles platform. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 44(4):1406–1419.

Jun Liu, Pu Zhao, Zhenglun Kong, Xuan Shen, Peiyan Dong, Fan Yang, Lin Cui, Hao Tang, Geng Yuan, Wei Niu, and 1 others. 2026f. When should a robot think? resource-aware reasoning via reinforcement learning for embodied robotic decision-making. *arXiv preprint arXiv:2603.16673*.

Xiaona Liu, Qing Wang, Minghao Zhou, Yanfei Wang, Xuefeng Wang, Xiaobo Zhou, and Qianqian Song. 2024b. Drugformer: Graph-enhanced language model to predict drug sensitivity. *Advanced Science*, 11(40):2405861.

Yinan Liu, Zihan Zhou, Zichun Jin, Xinyu Wang, Bin Wang, and Xiaochun Yang. 2026g. Enhancing llms in predictive political qa with semi-structured data. *Preprint*, arXiv:2608.21218.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025d. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. 2024c. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*.

Hanjun Luo, Shenyu Dai, Chiming Ni, Xinfeng Li, Guibin Zhang, Kun Wang, Tongliang Liu, and Hanan Salam. 2026a. Agentauditor: Human-level safety and security evaluation for llm agents. *Advances in Neural Information Processing Systems*, 38:43241–43298.

Hanjun Luo, Ziye Deng, Ruizhe Chen, and Zuozhu Liu. 2024a. Faintbench: A holistic and precise benchmark for bias evaluation in text-to-image models. *arXiv preprint arXiv:2405.17814*.

Hanjun Luo, Ziye Deng, Haoyu Huang, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. 2024b. Versusdebias: Universal zero-shot debiasing for text-to-image models via slm-based prompt engineering and generative adversary. *arXiv preprint arXiv:2407.19524*.

Hanjun Luo, Haoyu Huang, Ziye Deng, Xinfeng Li, Hewei Wang, Yingbin Jin, Yang Liu, Wenyuan Xu, and Zuozhu Liu. 2024c. Bigbench: A unified benchmark for evaluating multi-dimensional social biases in text-to-image models. *arXiv preprint arXiv:2407.15240*.

Hanjun Luo, Zhimu Huang, Sylvia Chung, Yiran Wang, Yingbin Jin, Jialin Li, Jiang Li, Xinfeng Li, and Hanan Salam. 2026b. Atelierevaluation of humans & llms as text-to-image prompters. *arXiv preprint arXiv:2605.22645*.

Hanjun Luo, Zhimu Huang, Haoyu Huang, Ziye Deng, Ruizhe Chen, Xinfeng Li, Zuozhu Liu, and Hanan Salam. 2026c. Biasig: Benchmarking multi-dimensional social biases in text-to-image models. *arXiv preprint arXiv:2604.11934*.

Hanjun Luo, Yingbin Jin, Yiran Wang, Xinfeng Li, Tong Shang, Xuecheng Liu, Ruizhe Chen, Kun Wang, Hanan Salam, Qingsong Wen, and 1 others. 2025a.

Dynamicner: A dynamic, multilingual, and fine-grained dataset for llm-based named entity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16522–16546.

Hanjun Luo, Chiming Ni, Jiaheng Wen, Zhimu Huang, Yiran Wang, Bingduo Liao, Sylvia Chung, Yingbin Jin, Xinfeng Li, Wenyuan Xu, and 1 others. 2025b. Centaureval: Benchmarking human-in-the-loop value in agentic coding. *arXiv preprint arXiv:2512.04111*.

Mikołaj Małkiński and Jacek Mańdziuk. 2023. A review of emerging research directions in abstract visual reasoning. *Information Fusion*, 91:713–736.

Mikołaj Małkiński and Jacek Mańdziuk. 2025. Deep learning methods for abstract visual reasoning: A survey on raven's progressive matrices. *ACM Computing Surveys*, 57(7):1–36.

Chaojun Meng, Changfan Pan, Hongji Shu, Qing Wang, Hanghui Guo, and Jia Zhu. 2025. Heterogeneous collaborative filtering contrastive learning for social recommendation. *Applied Soft Computing*, 173:112934.

Sumeet Ramesh Motwani, Alesia Ivanova, Ziyang Cai, Philip Torr, Riashat Islam, Shital Shah, Christian Schroeder de Witt, and Charles London. 2025. h1: Bootstrapping llms to reason over longer horizons via reinforcement learning. *arXiv preprint arXiv:2510.07312*.

OpenAI. 2025. Introducing o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/.

OpenAI. 2026. Gpt-5.4 thinking system card. https://deploymentsafety.openai.com/gpt-5-4-thinking/model-data-and-training.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mengting Pan, Fan Li, Chen Chen, and Xiaoyang Wang. 2026. Universal graph backdoor defense: A feature-based homophily perspective. *arXiv preprint arXiv:2605.16815*.

Mengting Pan, Fan Li, Chen Chen, Xiaoyang Wang, and Wenjie Zhang. 2025. Hitec: Hierarchical contrastive learning on text-attributed hypergraph with semantic-aware augmentation. *arXiv preprint arXiv:2508.03104*.

Tan-Hanh Pham and Chris Ngo. 2025. Rarl: Improving medical vlm reasoning and generalization with reinforcement learning and lora under data and hardware constraints. *arXiv preprint arXiv:2506.06600*.

Can Polat, Hasan Kurban, Erchin Serpedin, and Mustafa Kurban. 2025. Stress-testing multimodal foundation models for crystallographic reasoning. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 49–58.

Pin Qian, Su Wang, Xiaoyuan Wang, Yihang Chen, Wenxuan Xu, Qiaolin Yu, Shuhuai Lin, Sipeng Zhang, Junxian You, and Xinpeng Wei. 2026. Relevant is not warranted: Evidence-force calibration for cited rag. *arXiv preprint arXiv:2605.28044*.

Jialong Qin, Xin Zou, Di Lu, Yibo Yan, and Xuming Hu. 2026. Sharp eyes and memory for videollms: Information-aware visual token pruning for efficient and reliable videollm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8538–8546.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. 2025. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*.

Yilun Qiu, Jiahe Wang, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Chun Yuan. 2026. Agentcvr: Active multi-agent cross-video reasoning via script-simulated reinforcement learning. *arXiv preprint arXiv:2605.29643*.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. 2025. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6613–6629.

Yifan Shen, Chuanmiao Dong, Zhuoqing Zhong, Pei Tian, Tianjiao Yu, Jiateng Liu, Bowen Fang, Xinzhuo Li, Yuanzhe Liu, Zhengyuan Li, and 1 others. 2026a. Position: Multimodal llms should learn from children. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5619–5630.

Yifan Shen, Boyi Li, Meihuan Huang, Yuanzhe Liu, Anglin Liu, Junho Kim, Jingyuan Zhu, Fangzhou Lan, Jianguo Cao, Jintai Chen, Ismini Lourentzou, and James Matthew Rehg. 2026b. Decoding children's gait behavior. *arXiv preprint arXiv:2608.00371*.

Yifan Shen, Zhengyuan Li, and Gang Wang. 2024. Practical region-level attack against segment anything models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 194–203. IEEE.

Yifan Shen, Jiateng Liu, Xinzhuo Li, Yuanzhe Liu, Bingxuan Li, Houze Yang, Wenqi Jia, Yijiang Li, Tianjiao Yu, James Matthew Rehg, and 1 others. 2026c. Egoforge: Goal-directed egocentric world simulator. *arXiv preprint arXiv:2603.20169*.

Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. 2026d. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960.

Yifan Shen, Pei Tian, Xinzhuo Li, Bowen Fang, Shujun Xia, Bingxuan Li, Ana Jojic, Wenming Ye, Xu Cao, James Matthew Rehg, and 1 others. 2026e. Cogniroute: Learning to route social evidence in omnimodal models. *arXiv preprint arXiv:2606.20970*.

Yifan Shen, Jiawen Zhang, Jian Xu, Junho Kim, Ismini Lourentzou, Xu Cao, and Meihuan Huang. 2026f. Evaluating cognitive age alignment in interactive ai agents. *arXiv preprint arXiv:2605.17894*.

Hongji Shu, Chaojun Meng, Pasquale De Meo, Qing Wang, and Jia Zhu. 2024. Self-supervised hypergraph learning for enhanced multimodal representation. *IEEE Access*, 12:20830–20839.

Zitao Song, Yining Wang, Pin Qian, Sifan Song, Frans Coenen, Zhengyong Jiang, and Jionglong Su. 2023. From deterministic to stochastic: an interpretable stochastic model-free reinforcement learning framework for portfolio optimization. *Applied Intelligence*, 53(12):15188–15203.

Isabelle Soulières, Michelle Dawson, Fabienne Samson, Elise B Barbeau, Cherif P Sahyoun, Gary E Strangman, Thomas A Zeffiro, and Laurent Mottron. 2009. Enhanced visual processing contributes to matrix reasoning in autism. *Human brain mapping*, 30(12):4082–4107.

Guangyan Sun, Mingyu Jin, Zhenting Wang, Chenglong Wang, Siqi Ma, Qifan Wang, Tong Geng, Yingnian Wu, Yongfeng Zhang, and Dongfang Liu. 2025. Visual agents as fast and slow thinkers. In *International Conference on Learning Representations*, volume 2025, pages 6998–7023.

Xian Sun, Wei Gao, Yingshuo Wang, Lingdong Kong, Yanhang Li, Zhichao Fan, Zexin Zhuang, Wenlong Dong, Zhiyuan Zheng, Hrishikesh Paranjape, and 1 others. 2026. Beyond accuracy: Measuring bias acknowledgment in chain-of-thought reasoning for responsible ai evaluation. *arXiv preprint arXiv:2606.15127*.

Qitao Tan, Jun Liu, Zheng Zhan, Caiwen Ding, Yanzhi Wang, Xiaolong Ma, Jaewoo Lee, Jin Lu, and Geng Yuan. 2026. Harmony in divergence: Towards fast, accurate, and memory-efficient zeroth-order llm finetuning. *Advances in Neural Information Processing Systems*, 38:175003–175037.

Yicheng Tao, Yuanhao Huang, Yiqun Wang, Xin Luo, and Jie Liu. 2025. Autopcr: Automated phenotype concept recognition by prompting. In *arXiv preprint arXiv:2507.19315*.

Guiyao Tie, Xueyang Zhou, Tianhe Gu, Ruihang Zhang, Chaoran Hu, Sizhe Zhang, Mengqu Sun, Yan Zhang, Pan Zhou, and Lichao Sun. 2025. Mmmr: Benchmarking massive multi-modal reasoning tasks. *arXiv preprint arXiv:2505.16459*.

Can Wang, Yuhao Wang, Yushe Cao, Canran Xiao, and Fei Shen. 2026a. Lap-forensics: Latent-pixel consistency guided multimodal reasoning for deepfake detection. *arXiv preprint arXiv:2607.25962*.

Hexuan Wang, Yaxuan Ren, Srikar Bommireddypalli, Shuxian Chen, Adarsh Prabhudesai, Rongkun Zhou, Elina Baral, and Philipp Koehn. 2026b. Scitarc: Benchmarking qa on scientific tabular data that requires language reasoning and complex computation. *arXiv preprint arXiv:2603.08910*.

Hexuan Wang, Jingyu Zhang, Benjamin Van Durme, and Daniel Khashabi. 2026c. Are finer citations always better? rethinking granularity for attributed generation. *arXiv preprint arXiv:2604.01432*.

Jiaqi Wang, Kevin Qinghong Lin, James Cheng, and Mike Zheng Shou. 2026d. Think or not? selective reasoning via reinforcement learning for vision-language models. *Advances in Neural Information Processing Systems*, 38:111196–111234.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, and 1 others. 2026e. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*.

Qing Wang, Bo Li, Jialu Liang, Daling Shi, Bob Zhang, and Qianqian Song. 2026f. Drugclaw and drugaudit: A primary-source-grounded agent and authority-aware benchmark for drug-information question answering. *arXiv preprint arXiv:2606.01434*.

Qing Wang, Tianshi Liu, Minghao Zhou, Jialu Liang, Sen Guo, Guangyu Wang, Jing Su, and Qianqian Song. 2026g. Unid^3: A knowledge graph-enhanced rag framework for drug-disease discovery and reasoning. *arXiv preprint arXiv:2606.01394*.

Qing Wang, Jia Zhu, Hongji Shu, Kwame Omono Asamoah, Jianyang Shi, and Cong Zhou. 2023.

Gudn: a novel guide network with label reinforcement strategy for extreme multi-label text classification. *Journal of King Saud University Computer and Information Sciences*, 35(4):161–171.

Qirui Wang, Qi Guo, Yiding Sun, Junkai Yang, Dongxu Zhang, Shanmin Pang, and Qing Guo. 2026h. Personalq: Select, quantize, and serve personalized diffusion models for efficient inference. *arXiv preprint arXiv:2603.22943*.

Songrui Wang, Yubo Zhu, Wei Tong, and Sheng Zhong. 2024. Detecting dataset abuse in fine-tuning stable diffusion models for text-to-image synthesis. *arXiv preprint arXiv:2409.18897*.

Tongyan Wang, Zhengyuan Li, Muhan Lin, Shengyang Luo, Yifan Shen, Aniket Bera, Baijian Yang, and Yingjie Victor Chen. 2026i. Finemola: Towards fine-grained motion-language alignment from clip-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5510–5519.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2026j. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *Advances in Neural Information Processing Systems*, 38:118818–118850.

Yingshuo Wang, Xian Sun, Lingdong Kong, Wei Gao, Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026k. Do time series foundation model benchmarks hide regime-dependent failures? evidence from traffic speed forecasting. *arXiv preprint arXiv:2606.18367*.

Yingshuo Wang, Xian Sun, Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026l. Auditing and fixing economic validity in tabular foundation models for discrete choice. *arXiv preprint arXiv:2605.26559*.

Yingshuo Wang, Xian Sun, Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026m. Embedding foundation model predictions in discrete-choice models with structural guarantees. *arXiv preprint arXiv:2606.26432*.

Yining Wang, Zhixiang Lu, Pin Qian, Jionglong Su, Mian Zhou, Chong Li, and Zhengyong Jiang. 2025. Gems-llm: Integrating large language models with goal-aware exploration for rl-based portfolio optimization. In *International Conference on Intelligent Computing*, pages 516–527. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jiaqi Wei, Xuehang Guo, Pengfei Yu, Xiang Zhang, Wanli Ouyang, Siqi Sun, Qingyun Wang, and Chenyu You. 2026a. When to think, when to speak: Learning disclosure policies for llm reasoning. *arXiv preprint arXiv:2605.03314*.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, and 1 others. 2025a. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*.

Jiaqi Wei, Xiang Zhang, Yuejin Yang, Wenxuan Huang, Juntai Cao, Sheng Xu, Xiang Zhuang, Zhangyang Gao, Muhammad Abdul-Mageed, Laks VS Lakshmanan, and 1 others. 2025b. Unifying tree search algorithm and reward design for llm reasoning: A survey. *arXiv preprint arXiv:2510.09988*.

Jiaqi Wei, Hao Zhou, Xiang Zhang, Di Zhang, Zijie Qiu, Noah Wei, Jinzhe Li, Wanli Ouyang, and Siqi Sun. 2026b. Retrieval is not enough: Enhancing rag through test-time critique and optimization. *Advances in Neural Information Processing Systems*, 38:21484–21520.

Zhixiang Wei, Yi Li, Zhehan Kan, Xinghua Jiang, Zuwei Long, Shifeng Liu, Hongze Shen, Wei Liu, Xiaoyu Tan, Haojia Lin, and 1 others. 2026c. Youtu-vl: Unleashing visual potential via unified vision-language supervision. *arXiv preprint arXiv:2601.19798*.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*.

Shuai Xiao, Su Liu, Weikai Zhou, Jialun Wu, Xinjie He, Zhiyuan Lin, and Qiyang Xie. 2026. When does persona prompting actually help? a retrieval and metric analysis of expert role injection in llms. *arXiv preprint arXiv:2605.29420*.

Jingyi Xu, Tushar Vaidya, Yufei Wu, Saket Chandra, Zhangsheng Lai, and Kai Fong Ernest Chong. 2023. Abstract visual reasoning: An algebraic approach for solving raven's progressive matrices. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6715–6724. IEEE.

Junhao Xu, Zheng Liu, Mengqi Zhang, Guangyuan Dong, and Zhumin Chen. 2026. Lever can move the earth: Towards adaptive semantic capacity balance for image-text retrieval. *IEEE Transactions on Multimedia*.

Zhihan Yang, Jiaqi Wei, Xiang Zhang, Haoyu Dong, Yiwen Wang, Xiaoke Guo, Pengkun Zhang, Yiwei Xu, and Chenyu You. 2026. Forestllm: Large language models make random forest great on few-shot tabular learning. *arXiv preprint arXiv:2601.11311*.

Zhen Yao, Bo Lang, Zhihao Zheng, Haoyan Xu, Zhengtao Yao, Zhikang Dong, and Mooi Choo Chuah. 2026a. From semantic to reasoing: A survey of language-guided segmentation in the multimodal era.

Zhen Yao, Xin Li, Taotao Jing, Shuai Zhang, and Mooi Choo Chuah. 2026b. Disa: Saliency-aware foreground-background disentangled framework for

open-vocabulary semantic segmentation. *arXiv preprint arXiv:2601.20064*.

Zhen Yao, Xin Li, Bo Lang, and Mooi Choo Chuah. 2023. Goal-lbp: Goal-based local behavior guided trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 25(7):6770–6779.

Zhen Yao, Jiawei Xu, Shuhang Hou, and Mooi Choo Chuah. 2024. Cracknex: a few-shot low-light crack segmentation model based on retinex theory for uav inspections. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11155–11162. IEEE.

Chengzhen Yu, Canran Xiao, Siyuan Ma, and Yang Liu. 2026a. Text as partial constraint: Core-residual alignment for robust vision-language learning. *arXiv preprint arXiv:2607.03143*.

Haoran Yu, Xiaochong Jiang, Lifei Liu, Su Wang, Pin Qian, and Yihang Chen. 2026b. Beyond simpson's paradox: A cascade of confounders in ai agent pull-request co-authorship. *arXiv preprint arXiv:2606.22711*.

Haoran Yu, Lifei Liu, Xiaochong Jiang, Yuwen Jia, Su Wang, Pin Qian, and Yihang Chen. 2026c. Habituation at the gate: Rising approval and declining scrutiny in human review of ai agent code. *arXiv preprint arXiv:2606.22721*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2026d. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, and 1 others. 2025. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *arXiv preprint arXiv:2507.22607*.

Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. 2022. A benchmark for compositional visual reasoning. *Advances in neural information processing systems*, 35:29776–29788.

Daoan Zhang, Shuo Zhang, Zijian Jin, Jiebo Luo, Shengyu Fu, and Elsie Nallipogu. 2026a. Sphinx: Benchmarking and modeling for llm-driven pull request review. *arXiv preprint arXiv:2601.04252*.

Haichao Zhang, Mingfei Chen, and 1 others. 2026b. A survey of physical ai: A history from chatgpt to world models and embodied agents. *Preprints*.

Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. 2026c. Memmark: State-evolution attribution watermarking for agent long-term memory systems. *arXiv preprint arXiv:2605.25002*.

Huanyu Zhang, Wenshan Wu, Chengzu Li, Ning Shang, Yan Xia, Yangyu Huang, Yifan Zhang, Li Dong, Zhang Zhang, Liang Wang, and 1 others. 2025a. Latent sketchpad: Sketching visual thoughts to elicit multimodal reasoning in mllms. *arXiv preprint arXiv:2510.24514*.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2025b. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pages 68194–68213.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025c. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1859–1869. IEEE.

Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, and Yu Liu. 2023a. Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In *The Eleventh International Conference on Learning Representations*.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, and 1 others. 2026d. Theory of space: Can foundation models construct spatial beliefs through active exploration? *arXiv preprint arXiv:2602.07055*.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Shanghang Zhang, Gao Peng, and 1 others. 2025d. Mavis: Mathematical visual instruction tuning with an automatic data engine. In *International Conference on Learning Representations*, volume 2025, pages 87955–87989.

Xiang Zhang, Juntai Cao, Jiaqi Wei, Chenyu You, and Dujian Ding. 2025e. Why does your cot prompt (not) work? theoretical analysis of prompt space complexity, its interaction with answer space during cot reasoning with llms: A recurrent perspective. *arXiv preprint arXiv:2503.10084*.

Xinyu Zhang, Yuxuan Dong, Lingling Zhang, Chengyou Jia, Zhuohang Dang, Basura Fernando, Jun Liu, and Mike Zheng Shou. 2026e. Cofft: Chain of foresight-focus thought for visual language models. *Advances in Neural Information Processing Systems*, 38:106182–106201.

XiuYu Zhang, Yi Shan, Junfeng Fang, and Zhenkai Liang. 2026f. Self-evaluation is already there: Eliciting latent judge calibration in base llms with minimal data. *arXiv preprint arXiv:2606.05122*.

Yikai Zhang, Jiaxin Pei, Kenan Li, Qirui Jin, Maoquan Wang, Jin Pan, Yu Kang, Shengyu Fu, Elsie Nallipogu, Junjie Hu, and 1 others. 2026g. Swe-edit: Re-

thinking code editing for efficient swe-agent. *arXiv preprint arXiv:2604.26102*.

Yunfei Zhang, Yizhuo He, Yuanxun Shao, Zhengtao Yao, Haoyan Xu, Junhao Dong, Zhen Yao, and Zhikang Dong. 2026h. Chromouvqa: Benchmarking vision-language models under chromatic camouflaged images. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12777–12781. IEEE.

Yuzhe Zhang, Feiran Liu, Yi Shan, Xinyi Huang, Xin Yang, Yueqi Zhu, Xuxin Cheng, Cao Liu, Ke Zeng, Terry Jingchen Zhang, and 1 others. 2026i. Silobench: A scalable environment for evaluating distributed coordination in multi-agent llm systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29379–29398.

Zhilin Zhang, Xiang Zhang, Jiaqi Wei, Yiwei Xu, and Chenyu You. 2025f. Postergen: Aesthetic-aware paper-to-poster generation via multi-agent llms. *arXiv e-prints*, pages arXiv–2508.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ziliang Zhang, Yubo Zhu, Wei Tong, Jingyu Hua, Zijian Wang, Yuan Zhang, and Sheng Zhong. 2026j. Mitigating database leakage in rag systems with keyword-grounded fact substitution. *arXiv preprint arXiv:2608.21656*.

Chuangxin Zhao, Canran Xiao, Siyuan Ma, Mengyao Lyu, Yanbiao Ma, Jun Xia, Guiguang Ding, and Yang Liu. 2026a. Attention-spectrum regularization for replay-free continual multimodal llms. *arXiv preprint arXiv:2606.23063*.

Yang Zhao, Chengxiao Dai, Mengying Kou, and Yue Xiu. 2026b. Memorepair: Barrier-first cascade repair in agentic memory. *arXiv preprint arXiv:2605.07242*.

Yang Zhao, Chengxiao Dai, Yue Xiu, Mengying Kou, Yuliang Zheng, and Dusit Niyato. 2026c. Shardmemo: Masked moe routing for sharded agentic llm memory. *arXiv preprint arXiv:2601.21545*.

Yang Zhao, Chengxiao Dai, Wei Zhuo, Yue Xiu, and Dusit Niyato. 2026d. Clause: Agentic neuro-symbolic knowledge graph reasoning via dynamic learnable context engineering. In *International Conference on Learning Representations*, volume 2026, pages 82680–82706.

Yang Zhao, Yue Xiu, Chengxiao Dai, Ning Wei, and Dusit Niyato. 2025. Movable antenna enhanced federated fine-tuning of large language models via hybrid client selection optimization. *arXiv preprint arXiv:2506.00011*.

Zhihao Zheng, Zhen Yao, and Mooi Choo Chuah. 2026a. Luna-seg: Learning language-guided low-rank adaptation for night-time semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7586–7595.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. 2026b. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 126775–126798.

Liangyu Zhong, Fabio Philipp Rosenthal, Joachim Sicking, Fabian Hüger, Thorsten Bagdonat, Hanno Gottschalk, and Leo Schwinn. 2026. Focus: Internal mllm representations for efficient fine-grained visual question answering. *Advances in Neural Information Processing Systems*, 38:96498–96535.

Heng Zhou, Jing Tang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, Jiawei Yao, and 1 others. 2026a. Comem: Compositional concept-graph memory for vision–language adaptation. In *International Conference on Learning Representations*, volume 2026, pages 20009–20032.

Zihan Zhou, Yinan Liu, Yuyang Xie, Bin Wang, Xiaochun Yang, and Zezheng Feng. 2026b. Diaglink: A dual-user diagnostic assistance system by synergizing experts with llms and knowledge graphs. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Haokun Zhu, Zongtai Li, Zhixuan Liu, Wenshan Wang, Ji Zhang, Jonathan Francis, and Jean Oh. 2025a. Strive: Structured representation integrating vlm reasoning for efficient object navigation. *arXiv preprint arXiv:2505.06729*.

Yubo Zhu, Zhehan Kan, Jingyi Yang, Miaolin Chen, Jinbo Xing, Kai Zhu, Zijian Wang, Sheng Zhong, and Wei Tong. 2026a. When does visual generation help visual understanding in unified multimodal models? *arXiv preprint arXiv:2608.22174*.

Yubo Zhu, Dongrui Liu, Zecheng Lin, Wei Tong, Sheng Zhong, and Jing Shao. 2025b. The llm already knows: Estimating llm-perceived question difficulty via hidden representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1176.

Yubo Zhu, Songrui Wang, Wei Tong, Zijian Wang, and Sheng Zhong. 2026b. Protecting against unauthorized dataset use in fine-tuning text-to-image diffusion models. *IEEE Transactions on Information Forensics and Security*.

Xin Zou, Haolin Deng, Yibo Yan, Shuliang Liu, Zhiwei Jin, Chen Chen, Haonan Lu, and Xuming Hu. 2026a. Clearer sight, fewer lies: Oriented pickup preference optimization for multimodal hallucination mitigation. *arXiv preprint arXiv:2606.29805*.

Xin Zou, Haolin Deng, Yibo Yan, Shuliang Liu, Kening Zheng, Zhiwei Jin, Chen Chen, Haonan Lu, and Xuming Hu. 2026b. Consistency as inductive bias: Learning cross-view invariance for robust multimodal reasoning. *arXiv preprint arXiv:2606.29812*.

Xin Zou, Ruimeng Liu, Chang Tang, Zhenglai Li, Xinwang Liu, Kunlun He, and Wanqing Li. 2026c. Learning disentangled representations for generalized multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2026d. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

A Additional information of the Dataset

A.1 Statistics of the Dataset

Our Vbvr-VQA dataset is constructed by combining and processing the VBVR-Dataset and VBVR-Bench-Data from the Very Big Video Reasoning (VBVR) Suite (Wang et al., 2026e). Specifically, the VBVR-Dataset serves as the large-scale video reasoning training resource for the VBVR Suite. The original dataset is massive, containing 1,000,000 video clips generated by 100 distinct reasoning task generators (10,000 samples per generator). This results in approximately 5,000,000 individual files with a total size of roughly 370 GB. Furthermore, the VBVR-Bench-Data contains 500 test samples covering 100 tasks. Among these, 50 tasks are used for in-domain testing, and the remaining 50 are used for out-of-domain testing.

The data splits for the training and testing sets of Vbvr-VQA strictly follow the strategy of the original VBVR Suite. We use the frame ordering tasks processed from the VBVR-Dataset to train the ChronoVision framework. Subsequently, we evaluate the model on the tasks processed from the VBVR-Bench-Data. To ensure a rigorous evaluation, the training and testing tasks are generated using different random seeds.

The structure of our Vbvr-VQA is organized as follows.

Training Set: A total of 1,000,000 samples are used for the Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) stages of ChronoVision.

Testing Set: A total of 500 samples are used for the final benchmark evaluation. To rigorously test the model's generalization capabilities, this set is strictly divided into two parts: an **In-Domain Test Split** with 250 samples, where the task types and physical scenarios have appeared in the training set, and a **Zero-Shot Out-of-Domain (OOD) Test Split** with 250 samples. The OOD split contains completely new task generators and extreme physical scenarios that the model has never seen during the training phase, and is specifically designed to evaluate the generalized understanding of universal physical laws.

While the test set comprises 500 samples, this scale is strictly sufficient to demonstrate robust reasoning capabilities due to the mathematical formulation of the task and the depth of the physical scenarios. Mathematically, ordering six unordered candidate frames results in $6! = 720$ possible permu-

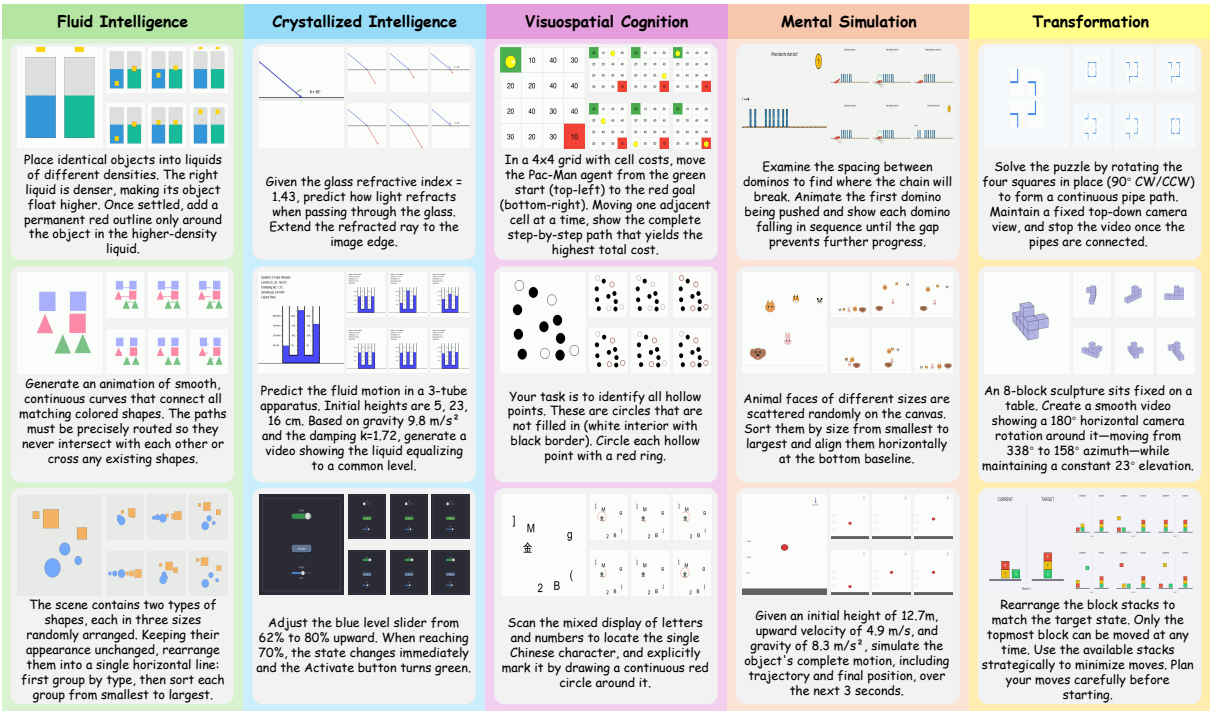

Figure 5: More Examples of Vbvr-VQA dataset

tations. Consequently, the probability of correctly guessing the exact chronological order by chance is exceptionally low ($1/720 \approx 0.14\%$). Achieving an accuracy exceeding 70 percent across 500 instances, particularly on the 250 out-of-domain samples, provides strong statistical significance. This proves that the model executes genuine temporal reasoning rather than relying on random selection or superficial statistical shortcuts.

Beyond statistical significance, the evaluation depth of each individual sample far exceeds standard visual question answering tasks. As illustrated in Figure 5, the test set covers 100 distinct reasoning task generators spanning complex cognitive domains such as fluid dynamics, spatial rotation, and kinematics. Each instance requires the model to perform multi-step mental simulation to reconstruct the visual evolution. The out-of-domain split introduces entirely unseen physical environments and extreme scenarios. This density of physical complexity justifies the test set scale, as the evaluation focuses on the rigorous depth of cognitive simulation rather than simple pattern recognition over massive data volumes.

A.2 Visual Examples of Dataset

To further illustrate the diversity of our Vbvr-VQA benchmark, we provide additional visual examples. As shown in Figure 1 (in the main text), our dataset covers five distinct cognitive categories: Fluid Intelligence, Crystallized Intelligence, Mental Simulation, Visuospatial Cognition, and Transformation. In Figure 5, we present another set of examples. These samples intuitively demonstrate the broad coverage, physical complexity, and high level of challenge in the tasks.

As mentioned in the main text, we reformulate the continuous video streams into a strict image ordering task. We uniformly segment the video along the temporal axis into six equal-length intervals and extract the final frame of each interval. These six discrete frames precisely capture the dynamic temporal evolution of the scene. Subsequently, we randomly shuffle these six images (labeled A through F) to serve as the candidate options for the model.

A.3 Annotation Details of Dataset

The text descriptions (prompts) are constructed by concatenating task instructions (requiring the model to order the images, such as "Please determine the correct chronological order of the following images based on physical plausibility and causal logic") with the original video descriptions provided by the VBVR dataset (which serve as prompts introducing the temporal logic). We use these combined prompts as the reasoning queries for the Multimodal Large Language Model (MLLM). During the inference process, the model

receives the initial frame, the text prompt, and the six shuffled candidate frames. This evaluates its ability to reconstruct the correct chronological order based on the given instructions.

Additionally, we use GPT-5 to generate dense spatio-temporal annotations. These annotations include not only precise bounding box coordinates for locating state change regions but also verb-centric semantic cues encapsulated in the <LOCATE>...</LOCATE> format. This provides high-quality supervision signals for our ROI Attention Locating module. We show our dataset pipeline in Figure 6 and related prompts in Figure 9.

A.4 Data Quality Assurance and Human Validation

A legitimate concern when utilizing data annotated by large language models is the potential risk of distilling hallucinated logic or erroneous physical assumptions. To prevent this, we strictly decouple the sequence determination from the text generation process. The chronological order of the candidate frames is an absolute ground truth, intrinsically defined by the continuous temporal axis of the original source video. GPT-5 is not tasked with inferring or guessing this sequence. Instead, the model is explicitly conditioned on the fixed ground-truth order. Its sole objective is to synthesize intermediate reasoning steps and extract spatial bounding box coordinates that explain the predetermined visual evolution.

To guarantee the validity of the evaluation benchmark, human annotators manually reviewed all 500 instances in the testing set. This review process verified that the generated temporal logic and semantic visual cues accurately reflect the actual physical events without ambiguity. Furthermore, to evaluate the reliability of the training data, we conducted a quality check by randomly sampling 500 instances from the training set. The manual inspection confirmed that the generated rationales logically align with the visual transformations. This rigorous verification confirms that the dataset provides physically sound supervision rather than flawed synthetic logic.

For further details on the ethics statement and reproducibility, please refer to Appendix G.

B More Implementation Details

In this section, we provide extended details regarding the training paradigm, hyperparameters, and the specific mechanisms of the Attention Condensation module that were omitted from the main text due to space constraints. Note that the Reconstructive Visual Head (RVH) operates in parallel to these mechanisms, as detailed in Section 4.2.

B.1 Training Paradigm and Hyperparameters

All training and inference processes are conducted on 8 NVIDIA H100 GPUs, utilizing DeepSpeed ZeRO-3 optimization to manage memory efficiency. The overall training pipeline is divided into two sequential stages: the Supervised Fine-Tuning (SFT) stage and the Reinforcement Learning (RL) stage.

SFT Stage. During the SFT stage, we optimize the model using the AdamW optimizer with a peak learning rate of 2×10^{-5}, a global batch size of 128, and a weight decay of 0.05. The model is trained for 3 epochs using a cosine learning rate decay schedule with a 3% linear warmup. The overall objective is

$$\begin{aligned}\mathcal{L}_{\text{total}} = &-\sum_{t=1}^{|Y|} \log q_\theta(Y_t \mid X, Y_{<t}) \\ &+ \beta \cdot \text{MSE}(\hat{h}_{\text{final}}, h_{\text{final}}) + \alpha \mathcal{L}_{AC}\end{aligned} \tag{8}$$

We strictly decouple the loss balancing: we set $\beta = 0.1$ for the RVH term and $\alpha = 0.1$ for the Attention Condensation loss.

RL Stage. For the subsequent RL stage utilizing Group Relative Policy Optimization (GRPO), we reduce the learning rate to 1×10^{-6} to ensure policy stability. For each input prompt, we sample a group size of $G = 8$ responses. The clipping parameter is set to $\epsilon = 0.2$, and the Kullback-Leibler (KL) divergence penalty coefficient is fixed at 0.01. The reward balancing weights are empirically set to $\omega_1 = 1.0$ for the outcome reward, $\omega_2 = 0.5$ for the latent-grounding reward, and $\omega_3 = 0.1$ for the unsupervised focus reward.

B.2 Justification for Attention Alignment over Bounding Box Regression

While ground-truth bounding box coordinates are available in the Vbvr-VQA dataset, we deliberately utilize them to supervise internal attention distributions rather than optimizing the model to directly output explicit coordinate values (e.g., via an Intersection over Union loss). This design choice is

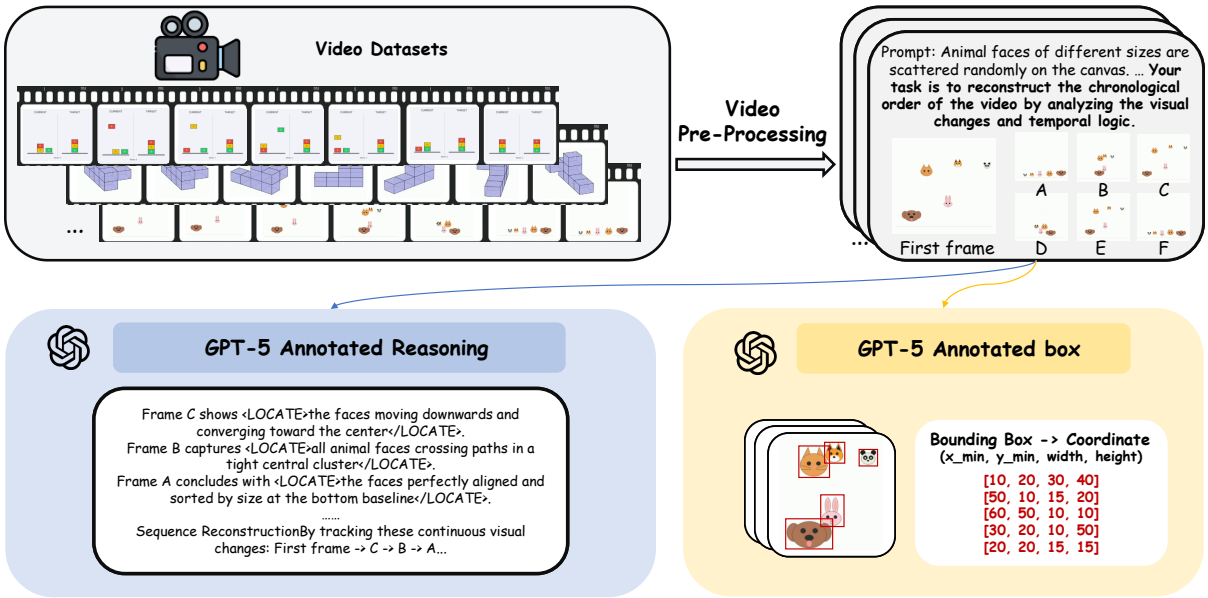

Figure 6: Dataset Processing Pipeline.

driven by the specific requirements of multi-step visual reasoning and the nature of physical simulation.

First, our core objective is to facilitate latent visual imagery. Explicitly predicting four discrete numerical coordinates compresses rich visual and spatial data into an extreme information bottleneck. By contrast, attention alignment preserves the high-dimensional, dense visual feature flow. These dense features are continuously processed by the Reconstructive Visual Head (RVH) to predict subsequent spatial states, maintaining feature integrity within the latent space.

Second, the Vbvr-VQA benchmark encompasses highly complex and non-rigid physical transformations, such as fluid dynamics, continuous smooth curves, and distributed objects. A rigid rectangular bounding box cannot accurately encapsulate irregular shapes or multiple scattered targets undergoing simultaneous state changes. Attention maps operate as pixel- or patch-level probability distributions, naturally accommodating arbitrary shapes and distributed regions without the structural constraints of explicit bounding boxes.

Finally, existing literature indicates that directly regressing continuous coordinate values using standard text-based Transformer architectures is often unstable and inefficient, frequently requiring specialized coordinate tokenizers. Supervising the self-attention mechanism via distribution alignment aligns directly with the fundamental architectural logic of multimodal large language models, ensuring stable spatial grounding.

B.3 Selection of the Designated Attention Layer

A core component of our methodology is condensing fragmented visual attention into a single designated intermediate layer ℓ. The selection of this layer ℓ is established prior to the fine-tuning phase via a heuristic approach, ensuring it does not overfit to specific downstream QA tasks.

To ensure generalizability, we evaluate the base (unfine-tuned) backbone on a held-out visual reasoning validation set. We extract the text-to-image self-attention weights across all layers and select the single layer that naturally exhibits the highest average attention concentration on the target bounding boxes $\mathcal{R}_{\text{tar}}$. Empirical observations demonstrate that the optimal layer ℓ consistently falls in the intermediate region of the Transformer blocks across different MLLM backbones. We freeze this layer selection for all subsequent training phases.

We explicitly opt for a *single* designated layer rather than aggregating multiple layers. Our preliminary experiments indicate that multi-layer aggregation dilutes the localization signal and introduces semantic noise from adjacent layers, whereas a single carefully selected intermediate layer provides the sharpest grounding interface for ROI extraction.

C Additional Analyses and Ablation Studies

C.1 Ablation Study on Evaluation with Partial Match Metrics

We select Exact Match as the primary metric because physical causality exhibits an all-or-nothing characteristic. In real-world continuous dynamics, such as fluid motion or object trajectories, reversing the chronological position of even two frames implies a fundamental break in the causal chain. Therefore, Exact Match serves as the strictest standard to verify whether the model fully comprehends the complete physical progression, rather than merely approximating visual states.

While Exact Match imposes a strict physical constraint, it assigns a score of zero to predictions with minor localized errors. To provide a more granular assessment of the models' ability to capture the general temporal trend, we evaluate the baseline models using Kendall's τ rank correlation coefficient as a partial match metric.

Kendall's τ measures the proportion of image pairs whose relative sequence is consistent between the predicted and ground-truth orders. For our task involving $N = 6$ frames, the total number of distinct frame pairs is $\binom{6}{2} = 15$. Let C denote the number of concordant pairs (where the predicted relative order matches the ground truth) and D denote the number of discordant pairs. Since there are no ties in our strict ordering format, $C + D = 15$. The metric is defined as:

$$\tau = \frac{C - D}{C + D} \tag{9}$$

This metric effectively isolates minor errors. Table 5 presents the Kendall's τ scores for ChronoVision and four prominent baselines across the ten cognitive categories. Even under this relaxed evaluation criterion, ChronoVision consistently demonstrates superior temporal alignment compared to the strong commercial and open-source models.

C.2 Ablation Study on Latent Sequence Evolution

To validate the effectiveness of the RVH and explore the evolution if the latent sequence generated by it, we track the trajectories of the predicted latent sequence during the reasoning process for correct and incorrect predicted image orderings. Specifically, we measure the MSE between the predicted latent sequence $\hat{h}_{\text{final}}$ generated by the RVH at various steps of the reasoning chain and the ground-truth sequence embedding of the final state h_{final} extracted by the frozen visual encoder. The results are averaged over 10 sampled reasoning trajectories. As shown in Table 6, for correctly ordered trajectories, the MSE decreases as the reasoning step increases, providing evidence that the RVH successfully guides the model to mentally simulate and converge toward the correct final ordering sequence step by step. For incorrect trajectories, the MSE fails to converge to the ground-truth latent sequence.

C.3 Ablation Study on Chain-of-Thought

To validate that the latent sequence reconstruction relies on step-by-step textual reasoning, we conduct an ablation study removing the intermediate reasoning traces, forcing the model to generate the final image ordering directly in a single-step reply. As demonstrated in Table 7, removing the CoT leads to a performance drop, demonstrating that the intermediate textual reasoning is crucial to guide and refine the predicted latent sequence $\hat{h}_{\text{final}}$ toward the correct image ordering.

C.4 Ablation Study on Reasoning Dependence

To directly test whether the RVH is merely passively following the textual CoT, we perform an ablation study in which we inject an incorrect intermediate reasoning prefix and force the model to continue generation from that step. As shown in 8, if the RVH only acted as a passive auxiliary signal, then the model should largely inherit the erroneous reasoning and collapse similarly to the SFT *w/o* head variant. Instead, we observe a much smaller degradation for the full ChronoVision model (SFT with ROI + RL) compared to the baseline. This indicates our training framework internalizes visually grounded constraints into the learned policy, making the model more robust to wrong intermediate reasoning prefixes.

C.5 Ablation Study on Intermediate State Perturbation

Following the established methodology (Li et al., 2026n), we test whether each generated latent sequence is necessary by selecting successful reasoning trajectories and injecting random noise into the latent sequence at a specific intermediate step. The model then continues decoding from the perturbed state. We measure the MSE between the predicted intermediate latent sequence $\hat{h}_{\text{final}}$ and

Models	In-Domain by Category					Out-of-Domain by Category				
	Flu.	Cry.	Vis.	Men.	Trans.	Flu.	Cry.	Vis.	Men.	Trans.
Qwen 3.5 9B	29.5	31.4	36.8	51.2	30.9	31.6	36.2	25.1	30.7	29.4
Gemini 3 Flash	55.6	68.0	69.6	84.4	62.3	55.9	59.4	64.0	65.8	71.4
Gemini 3.1 Pro	56.1	78.4	78.8	73.8	65.8	63.0	76.2	66.6	66.9	24.8
Claude Opus 4.6	58.8	80.1	76.2	74.1	69.9	70.1	81.0	66.4	74.2	83.5
ChronoVision	**88.4**	**90.7**	**89.5**	**87.8**	**91.0**	**86.9**	**89.5**	**88.8**	**85.5**	**89.3**

Table 5: Partial match evaluation using Kendall's τ. We report the metric across all In-Domain and Out-of-Domain categories. Overall and average metrics are omitted to focus on specific cognitive dimensions.

Table 6: Latent sequence evolution across reasoning steps. We report the MSE between the predicted final-state latent representation and the ground-truth final-state representation for correct and incorrect ordering trajectories.

Final Answer	Step 0	Step 10	Step 20	Step 30	Step 40
Correct Answer	0.512	0.364	0.245	0.182	0.146
Incorrect Answer	0.508	0.456	0.412	0.435	0.428

the final ground-truth latent sequence h_{final} at different intermediate reasoning steps. The results are averaged over 10 sampled reasoning trajectories. As shown in Table 9, the trajectory remains stable before the intervention. Once the state is perturbed, downstream latent error increases, and the trajectory fails to reconverge to the clean path. This provides evidence that intermediate latent sequences play a causal role in sustaining the reasoning trajectory, rather than merely being correlated with outcome confidence.

C.6 Ablation Study on Latent Sequence Intervention

To further test whether the RVH actively guides reasoning, we conduct an ablation study with latent patching experiment (Liang and Pan, 2026). We intervene during intermediate decoding by swapping the predicted latent sequence between correct and incorrect reasoning runs. We measure the change in the probability of generating the correct frame permutation after the intervention ($\Delta p = p_{\text{after}} - p_{\text{before}}$). As shown in Table 10, when we patch a distractor's latent sequence into a correct trajectory, the downstream rationale shifts toward that disaractor's reasoning trajectory with wrong image order. Patching a correct latent sequence into an incorrect run actively shifts the reasoning trajectory toward the correct sequence. This provides a strong evidence that the patched intermediate latent sequence carries vital causal information and actively guides the inference of visual transformations.

C.7 Ablation Study on Linear Probing

We also measure how predictive the intermediate latent sequence is of the final sequence permutation at different reasoning steps using a linear probe (Alain and Bengio, 2016). Specifically, we freeze ChronoVision and train a lightweight linear classifier to predict the exact image order only using the intermediate latent sequence $\hat{h}_{\text{final}}$ extracted at different steps of the reasoning process. If the model are taking a shortcut, the latent state would immediately identify the correct order from the start. As shown in Table 11, the accuracy increases gradually and progressively rather than appearing fully formed at early steps. This rules out the shortcut learning hypothesis, providing evidence that the target image order logic is truly constructed step-by-step via latent sequences as the textual reasoning unfolds.

C.8 Comparison with Related Temporal Reasoning Methods

We compare ChronoVision with R1-VL(Zhang et al., 2025c), VL-Cogito(Yuan et al., 2025), and Latent Sketchpad(Zhang et al., 2025a) using the same Qwen3.5-9B backbone and identical SFT data. ChronoVision obtains the highest overall, ID, and OOD accuracy. R1-VL and VL-Cogito supervise the reasoning process without an explicit target for the final visual state. Latent Sketchpad introduces visual latent representations, but its objective is not tied to the final state that determines the correct ordering. The results suggest that direct supervision of the final-state representation provides useful information for temporal ordering.

Table 7: **CoT Ablation Study Results.** We compare the performance of ChronoVision with and without CoT across the In-Domain categories of Vbvr-VQA benchmark. Forcing the model to generate the final sequence directly in a single step leads to a consistent performance drop, demonstrating that step-by-step textual reasoning is crucial for guiding the latent sequence generation.

Variant	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Transformation	Avg.
ChronoVision without CoT	66.2	71.1	68.3	60.0	69.1	67.6
ChronoVision with CoT	**72.3**	**77.8**	**75.0**	**68.0**	**78.2**	**74.8**

Table 8: **Reasoning Dependence Ablation Study Results.** We compare the SFT *w/o* head variant and full ChronoVision (SFT with ROI + RL) under standard inference and when forced to generate from an injected incorrect reasoning prefix. The full model exhibits less performance degradation, demonstrating that our framework internalizes visually grounded constraints to resist incorrect intermediate textual logic.

Variant	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Transformation	Avg.
SFT *w/o* head (forced incorrect prefix)	50.8	57.8	50.0	40.0	50.9	50.8
SFT *w/o* head (standard inference)	66.2	73.3	66.7	56.0	67.3	66.8
ChronoVision (forced incorrect prefix)	60.0	66.7	63.3	56.0	67.3	63.2
ChronoVision (standard inference)	72.3	77.8	75.0	68.0	78.2	74.8

Table 9: **Intermediate State Perturbation Ablation Study Results.** We report the MSE between the predicted intermediate latent sequence ($\hat{h}_{\text{final}}$) and the ground-truth final latent sequence (h_{final}) across different reasoning steps. Injecting Gaussian noise at Step 30 disrupts an otherwise successful reasoning trajectory. The perturbed trajectory diverges and fails to reconverge to the clean path. This demonstrates that intermediate latent sequences causally drive and sustain accurate reasoning.

Intervention	Step 0	Step 10	Step 20	Step 30	Step 40
Add Gaussian Noise @ Step 30	0.512	0.364	0.245	0.425	0.468
ChronoVision *w/o* Noise	0.512	0.364	0.245	0.182	0.146

D Additional Evaluation on Real-World Video Reasoning

D.1 Evaluation on Video-Holmes

We evaluate ChronoVision on Video-Holmes(Cheng et al., 2025) to test whether the learned temporal reasoning ability transfers beyond the Vbvr-VQA ordering setting. We uniformly sample 32 frames from each video for ChronoVision. Table 13 reports the results across seven reasoning categories. ChronoVision obtains an overall score of 45.89, compared with 42.00 for GPT-4o and 45.00 for Gemini-2.5-Pro. The result indicates that the model can transfer to real-world video reasoning tasks.

D.2 Evaluation on LongVideo-Reason

We also evaluate ChronoVision on LongVideo-Reason(Chen et al., 2026e), which measures temporal, goal, plot, and spatial reasoning over long videos. As shown in Table 14, ChronoVision reaches an overall score of 74.7. It also obtains the highest score in each of the four reported categories.

E Visualization Example

Figure 7 presents a representative mental simulation example from Vbvr-VQA. In this task, the model is required to recover the correct temporal order of six shuffled candidate frames, where a set of animal faces gradually transitions from a scattered initial layout to a horizontally aligned arrangement sorted by size along the bottom baseline. This example highlights that solving the task requires more than static frame matching: the model must track continuous positional changes, intermediate reorganization patterns, and the gradual formation of the target configuration. Compared with strong baselines, ChronoVision better captures this progressive visual evolution and reconstructs the correct sequence by grounding its reasoning in continuous visual changes.

Figure 8 shows a representative transformation example from Vbvr-VQA. Here, the task is to reconstruct the temporal order of a smooth camera orbit around a fixed 8-block sculpture, starting from the initial viewpoint and ending at the reversed orientation after a 180-degree horizontal rotation. Unlike the previous example, this problem mainly tests whether the model can track viewpoint-induced structural changes, including foreshorten-

Table 10: **Latent Sequence Intervention Ablation Study Results.** We report the change in probability ($\Delta p = p_{\text{after}} - p_{\text{before}}$) of generating the correct image ordering after intervening on the predicted latent sequence. Patching an incorrect distractor's latent sequence into a correct reasoning trajectory disrupts the generation, while patching a correct latent sequence into an incorrect run successfully guides the trajectory back toward the correct image order.

Intervention Direction (Source → Destination)	Fluid Intel.	Crystallized Intel.	Visuospatial Cog.	Mental Simu.	Transformation	Avg.
Incorrect → Correct Patch	-0.42	-0.38	-0.45	-0.40	-0.52	-0.44
Correct → Incorrect Patch	+0.25	+0.22	+0.30	+0.26	+0.35	+0.28

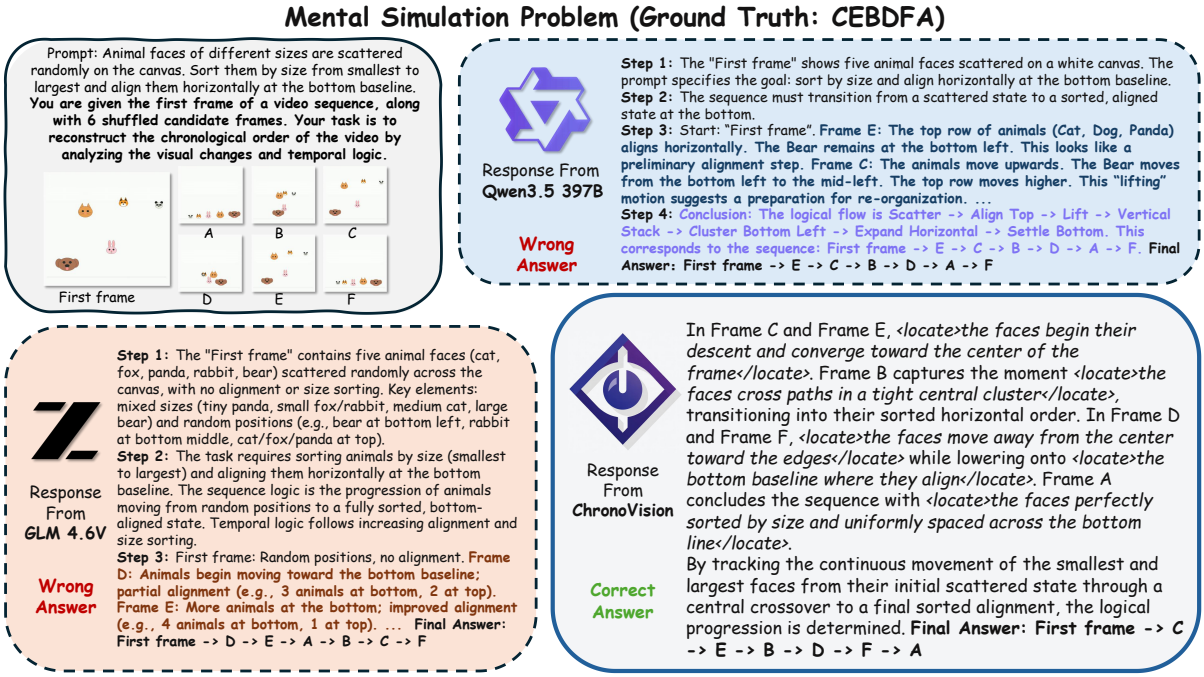

Figure 7: Qualitative comparison on a representative mental simulation problem from Vbvr-VQA. ChronoVision correctly reconstructs the temporal order by grounding reasoning in continuous visual changes.

Table 11: **Linear Probing Ablation Study Results.** We report the accuracy of a linear probe trained to predict the correct image order using only the intermediate latent sequence ($\hat{h}_{\text{final}}$) extracted at different reasoning steps. The probing accuracy on the In-Domain set increases as the reasoning chain unfolds, demonstrating that the temporal logic is constructed step-by-step rather than memorized via a shortcut.

	Step 0	Step 10	Step 20	Step 30	Step 40
ID Acc. (%)	18.4	31.6	49.2	65.5	74.1

ing, side-face visibility, and the gradual emergence of the opposite side of the object. The qualitative comparison shows that ChronoVision more accurately follows these continuous structural changes and therefore recovers the correct chronological order, whereas baseline models fail to maintain a consistent transformation path.

F Related Works

F.1 VLMs for Vision Reasoning

Recent advancements in Vision-Language Models (VLMs) have shifted focus from basic perceptual tasks toward complex deductive reasoning (Ke et al., 2025; Zhang et al., 2026e; Izadi et al., 2026; Zou et al., 2026b; Qin et al., 2026; Xu et al., 2026). To overcome the text bottleneck, where high-dimensional visual data is compressed into discrete tokens—recent frameworks propose reasoning directly within continuous visual and latent spaces, preserving crucial spatial cues while mitigating hallucinations and over-reliance on superficial language shortcuts (Qin et al., 2025; Zhang et al., 2025a; Li et al., 2026a, 2025d; Zhao et al., 2026a; Chen et al., 2026b; Yu et al., 2026a).

Aligning VLMs for protracted reasoning relies heavily on robust Reinforcement Fine-Tuning, notably Group Relative Policy Optimization (GRPO) (Ouyang et al., 2022; Shao et al., 2024).

Method	Overall	In-Domain						Out-of-Domain					
		Avg.	Flu.	Cry.	Vis.	Men.	Trans.	Avg.	Flu.	Cry.	Vis.	Men.	Trans.
SFT + R1-VL	70.4	72.0	70.8	75.6	71.7	64.0	74.5	68.8	64.4	72.5	70.0	60.0	71.0
SFT + VL-Cogito	70.8	72.4	70.8	75.6	73.3	64.0	74.5	69.2	64.4	72.5	70.0	60.0	72.0
SFT + Latent Sketchpad	69.8	71.2	69.2	75.6	71.7	64.0	72.7	68.4	64.4	70.0	70.0	60.0	71.0
ChronoVision	**73.2**	**74.8**	**72.3**	**77.8**	**75.0**	**68.0**	**78.2**	**71.6**	**66.7**	**75.0**	**72.5**	**64.0**	**74.0**

Table 12: Comparison with related temporal reasoning methods using the same Qwen3.5-9B backbone and identical SFT data. Accuracy is reported in percentage points. Flu., Cry., Vis., Men., and Trans. denote Fluid Intelligence, Crystallized Intelligence, Visuospatial Cognition, Mental Simulation, and Transformation.

Model	Frames	SR	IMC	TCI	TA	MHR	PAR	CTI	Overall
Qwen2.5-VL-7B	32	38.40	34.80	17.60	30.00	27.10	18.60	25.20	27.80
Video-R1	32	48.60	41.70	28.90	34.50	31.00	33.50	35.90	36.50
GPT-4o	32	50.00	49.60	38.80	30.00	44.00	39.20	37.00	42.00
Gemini-2.5-Pro	N/A	46.60	49.30	46.90	**53.00**	40.10	44.30	37.40	45.00
ChronoVision	32	48.97	48.55	**47.99**	52.00	**41.87**	**44.85**	**38.89**	**45.89**

Table 13: Results on Video-Holmes. We report performance across the seven reasoning categories and the overall benchmark score. ChronoVision uses 32 uniformly sampled frames for each video.

Recent works significantly extend this alignment through step-wise rewards, iterative SFT-RL cycles, and progressive curricula to refine complex logic and mitigate reasoning shortcuts (Zhang et al., 2025c; Deng et al., 2026; Xia et al., 2025; Yuan et al., 2025; Liu et al., 2026b; Zhou et al., 2026a; Liao et al., 2025). Furthermore, incorporating fine-grained visual token perception, selective Monte Carlo Tree Search, and self-reflective reasoning traces has proven to drastically enhance data efficiency and spatial grounding during RL optimization (Huang et al., 2026; Wang et al., 2026j,d; Chen et al., 2025a; Shen et al., 2026d).

To bridge perception and logic, visual instruction tuning explicitly aligns visual features with linguistic reasoning (Liu et al., 2023; Huang et al., 2025), driving progress across general multimodal benchmarks (Zhang et al., 2025d) and specialized constrained domains (Zhu et al., 2025a; Pham and Ngo, 2025). Recent unified multimodal models further explore how joint vision-language supervision and autoregressive modeling can improve visual understanding, while also examining when visual generation provides complementary benefits to perception and reasoning (Zhu et al., 2026a; Wei et al., 2026c; Kan et al., 2026). However, extending these reasoning capabilities to dynamic temporal contexts reveals a persistent gap between static spatial knowledge and true temporal understanding (Feng et al., 2025; Cores et al., 2024). For instance, recent studies on fine-grained action recognition demonstrate that state-of-the-art MLLMs still struggle to decode subtle temporal dynamics, such as pediatric gait behaviors from continuous video sequences (Shen et al., 2026b). Recent efforts have sought to bridge this gap by enforcing fine-grained motion-language alignment from video clips (Wang et al., 2026i) and modeling physical dynamics within embodied world simulators (Zhang et al., 2026b; Shen et al., 2026c). Recent work also explores agentic formulations of video reasoning, where multiple agents coordinate over cross-video evidence for complex temporal reasoning (Qiu et al., 2026). To address this temporal bottleneck, our work reformulates video reasoning as an image ordering task to enforce the understanding of visual evolution over time, further grounding the reasoning process by supervising the latent visual representation of the correct final state.

F.2 Visual Cognition

Vision-Language Models (VLMs) exhibit profound deficits in core knowledge, intuitive physics, and abstract reasoning compared to human intelligence (Li et al., 2024; Schulze Buschoff et al., 2025; Zou et al., 2024; Liu et al., 2026d; Zou et al., 2026a; Li et al., 2026j). This profound discrepancy is fundamentally rooted in the lack of developmental cognitive alignment; unlike human cognition which organically builds core physical concepts during childhood, current VLMs struggle with intuitive physics and require explicit alignment with

Model	Temporal	Goal	Plot	Spatial	Overall
LongVILA-7B	58.0	80.2	57.1	46.7	62.7
Video-R1-7B	61.4	85.0	62.0	58.5	68.1
Gemini-1.5-Pro	65.4	81.9	67.8	53.3	69.3
LongVILA-R1-7B	68.1	85.7	70.6	53.3	72.0
ChronoVision	**72.2**	**87.0**	**71.9**	**59.3**	**74.7**

Table 14: Results on LongVideo-Reason. We report temporal, goal, plot, spatial, and overall benchmark scores.

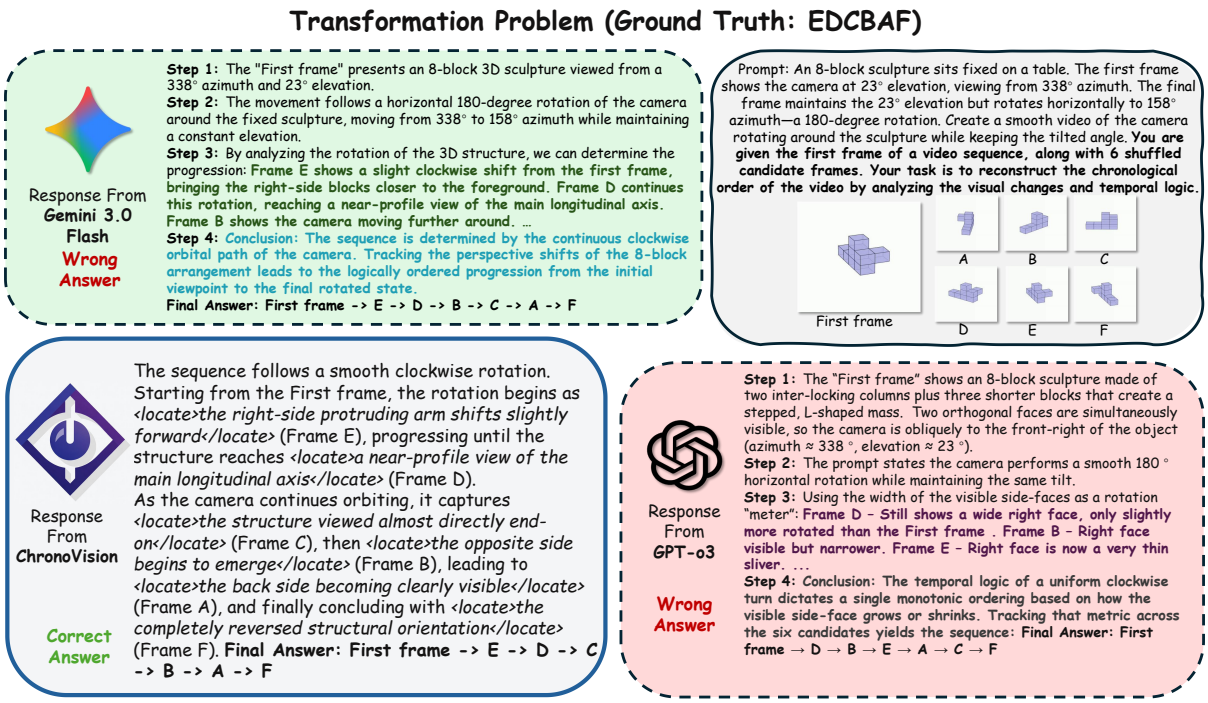

Figure 8: Qualitative comparison on a representative transformation problem from Vbvr-VQA. ChronoVision correctly reconstructs the temporal order by grounding reasoning in continuous structural changes during camera rotation.

human cognitive developmental stages (Shen et al., 2026a,f; Chen et al., 2026c; Zhang et al., 2026h; Wang et al., 2026a). Recent studies further highlight a substantial visual cognition gap between humans and multimodal LLMs in abstract visual reasoning (Cao et al., 2024; Li et al., 2026d). Benchmarks such as MaRs-VQA (Cao et al., 2024) and VRIQ (Khezresmaeilzadeh et al., 2026) reveal that current VLMs are fundamentally bottlenecked by perceptual limitations and multi-image relational cognition rather than language logic alone. Furthermore, comprehensive spatiotemporal assessments including IntPhys 2 (Bordes et al., 2025), OmniSpatial (Jia et al., 2026), Mmmr (Tie et al., 2025), and xCrysAlloys (Polat et al., 2025) demonstrate that VLMs consistently fail at object permanence, spatial dynamics, and physical consistency. To address these deficits, CogSense-Bench introduces a testbed spanning five cognitive dimensions (Li et al., 2026d). The Cognitive Supersensing paradigm overcomes text-only reasoning constraints by integrating latent visual imagery prediction, significantly bridging the gap between superficial perception and genuine visual cognition (Li et al., 2026d).

Matrix reasoning has long served as a core paradigm for studying abstract visual reasoning and its connection to human cognition (John and Raven, 2003; Soulières et al., 2009). Early work showed that neural models can capture certain compositional visual relations and solve simplified matrix reasoning problems (Fleuret et al., 2011; Małkiński and Mańdziuk, 2025, 2023; Xu et al., 2023). Building on this line, benchmark development has played a central role in advancing the field, including RAVEN (John and Raven, 2003), I-RAVEN (Hu et al., 2021), and CVR (Zerroug et al., 2022), which provide structured testbeds for eval-

uating relational and analogical reasoning. More recently, RAVEN-IQ (Huang et al., 2023) extends this direction by emphasizing zero-shot visual reasoning settings. By comparison, our work targets this visual cognition gap in temporal reasoning by explicitly training the model to internally simulate the latent visual state of the correct final outcome, rather than inferring it only through textual or symbolic shortcuts.

F.3 ROI Selection and Cropping

In visual reasoning tasks, the evidence required to answer a question is often confined to small yet informative regions, while VLMs typically process the entire image at limited resolution, making it difficult to capture dense text, subtle attributes, and fine-grained spatial cues. To address this issue, early studies mainly relied on tool-based ROI selection. Chain-of-Spot (Liu et al., 2024c) introduces localized region exploration into multi-step visual reasoning, while ZoomEye (Shen et al., 2025) further performs progressive zooming and region search to acquire key evidence, showing that active local cropping can substantially improve evidence acquisition in complex reasoning. Subsequently, Visual Agents as Fast and Slow Thinkers (Sun et al., 2025) and Refocus (Fu et al., 2025) incorporate supervised fine-tuning into ROI localization, enhancing local evidence modeling through step-wise decision making and explicit refocusing, respectively. More recent works such as DeepEyes (Zheng et al., 2026b) and DeepEyesV2 (Hong et al., 2026) further optimize multi-step region exploration with reinforcement learning, enabling models to actively acquire more informative visual evidence during inference. Beyond merely locating regions of interest, it is equally critical to ensure the robustness of these extracted features against pixel-grounding hallucinations (Cao et al., 2026; Li et al., 2026l; Shen et al., 2024; Yao et al., 2026b,a; Zheng et al., 2026a; Zou et al., 2026d; Zhang et al., 2023a) and to dynamically route Omni-modal evidence to avoid attention dispersion during multi-step inference (Shen et al., 2026e).

To avoid the brittleness of explicit coordinate prediction, later work shifted toward attention-driven ROI localization. Methods such as ICoT (Gao et al., 2025), ViCrop (Zhang et al., 2025b), and FOCUS (Zhong et al., 2026) derive relevant regions directly from internal attention maps rather than autoregressively generating bounding boxes, suggesting that ROI localization is closely tied to the model's internal "where-to-look" signals. However, such signals remain sensitive to layer-wise attention dispersion and to the textual query used for extraction, making stable localization still challenging. Unlike these approaches, our method uses semantically guided ROI attention locating to concentrate internal attention on dynamic key regions, thereby supporting the modeling of temporal visual evolution.

F.4 Reinforcement Learning

Reinforcement learning has become an effective paradigm for improving complex reasoning beyond supervised imitation. Recent work has advanced this training pipeline from multiple perspectives, including optimization stability, sampling efficiency, and long-horizon credit assignment. For example, DAPO (Yu et al., 2026d) stabilizes long-chain reasoning training through decoupled clipping and dynamic sampling, CPPO (Lin et al., 2026) improves efficiency by pruning low-contribution completions, and VinePPO (Kazemnejad et al., 2024) improves credit assignment through more accurate return estimation. Recent work has also systematically examined the interaction between search strategies and reward design for LLM reasoning (Wei et al., 2025b; Liu et al., 2026f).

On the optimization side, PPO (Schulman et al., 2017) remains a standard backbone for post-training, while GRPO (Shao et al., 2024) has become a practical choice for reasoning-oriented RL by estimating relative advantages from grouped sampled responses without training a separate value model. Recent analyses further identify biases in GRPO, such as a tendency toward longer responses, motivating refined variants such as Dr. GRPO (Liu et al., 2025d). Unlike prior work that mainly optimizes textual reasoning quality or answer-level outcomes, we adopt GRPO for implicit process grounding in temporal visual reasoning, using a composite reward over sequence correctness, latent visual alignment, and attention focus.

F.5 Foundation Model Reliability and Agents

Beyond multimodal reasoning, recent research has examined foundation models from several complementary perspectives. One line focuses on robustness, structural validity, and reliable evaluation across domain-specific or adversarial settings (Jiang et al., 2026; Wang et al., 2026m,l,k; Pan et al., 2026). Complementary work investigates representation learning and model adaptation

across generative, multi-view, graph-based, multimodal, and tabular settings (Wang et al., 2026h; Zou et al., 2026c; Pan et al., 2025; Zhao et al., 2025; Yang et al., 2026; Shu et al., 2024; Meng et al., 2025; Liu et al., 2024a, 2025a; Tan et al., 2026). For language models, substantial attention has been devoted to memory, prompting, chain-of-thought behavior, self-evaluation, and reasoning control (He et al., 2026; Xiao et al., 2026; Sun et al., 2026; Zhu et al., 2025b; Li et al., 2026e; Wei et al., 2026a; Zhang et al., 2025e, 2026f). Related studies investigate structured and knowledge-intensive reasoning, including graph-based inference, complex QA, and knowledge-boundary expansion (Li et al., 2025c; Zhang et al., 2026j; Li et al., 2025b, 2026k; Zhao et al., 2026d; Dai et al., 2025; Wang et al., 2026b). Retrieval grounding and attribution form another important reliability axis, spanning retrieval conflicts, test-time critique, reranking, evidence calibration, and citation quality (Chen et al., 2026d; Qian et al., 2026; Wei et al., 2026b; Li et al., 2026i; Wang et al., 2026c,g). Generative multimodal systems have similarly motivated extensive study of evaluation, bias, and trustworthy generation (Luo et al., 2026b, 2024c, 2026c, 2024a,b), while broader language applications explore dynamic information extraction, prompting, classification, generation, and graph-enhanced prediction (Tao et al., 2025; Luo et al., 2025a; Wang et al., 2023; Guo et al., 2026; Liu et al., 2024b).

A parallel body of work extends foundation models toward increasingly autonomous and stateful agents. Software-engineering agents have been studied for code maintenance, editing, repository execution, pull-request review, and the use of external information during problem solving (Jin et al., 2026a; Zhang et al., 2026g; Li et al., 2026g,h; Zhang et al., 2026a; Liu et al., 2026e, 2025b,c). Human–agent interaction and agent evaluation further examine collaboration, oversight, safety, and user adaptation (Yu et al., 2026c,b; Luo et al., 2025b, 2026a; Li et al., 2026f). Research and domain-specific agents broaden this paradigm toward autonomous scientific workflows, agent-native research, multi-agent generation, and grounded question answering (Liu et al., 2026c; Wei et al., 2025a; Zhang et al., 2025f; Wang et al., 2026f; Zhu et al., 2026b; Wang et al., 2024). Multi-agent coordination has also emerged as an important capability in distributed and dynamic environments (Zhang et al., 2026i; Dai et al., 2026a; Chen et al., 2024, 2026a, 2025b; Zhou et al., 2026b; Liu et al., 2026g; Jin et al., 2026b). Long-term memory constitutes another central component of agentic systems, motivating work on memory representation, attribution, routing, and repair (Liu et al., 2026a; Zhang et al., 2026c; Zhao et al., 2026c,b). Beyond language-centric agents, related research explores reinforcement-learning-based decision making, risk-aware intervention, and autonomous perception and control (Song et al., 2023; Wang et al., 2025; Dai et al., 2026b; Yao et al., 2023, 2024). Together, these directions reflect the growing emphasis on AI systems that must integrate reliable reasoning, adaptation, memory, coordination, and decision making in complex environments.

G Ethics Statement and Reproducibility

G.1 Artifact License and Data Usage

The Vbvr-VQA benchmark introduced in this paper is constructed by sampling and processing data from the open-source Very Big Video Reasoning (VBVR) dataset (Wang et al., 2026e). The research focus and applications of the two are essentially the same, and our use of VBVR is consistent with its intended use. The original VBVR dataset is publicly released under the Apache License 2.0. This license permits free use, modification, and distribution for academic and commercial purposes. Our usage, modifications, and the subsequent release of the Vbvr-VQA task splits strictly comply with the terms of the Apache License 2.0.

G.2 Human Validation Subjects and Instructions

To verify the physical correctness of the ground-truth sequences and the GPT-5 generated reasoning traces, we conducted human validation on the entire testing set (500 samples) and a random subset of the training set (500 samples).

Recruitment and Compensation. The validation process was performed by graduate-level researchers specializing in computer vision and physics-based machine learning. Since the validation was conducted as an internal academic review procedure to ensure data quality for the benchmark, it did not involve external crowdworkers. Consequently, the reviewers participated voluntarily as part of their academic research duties, and no direct monetary compensation was provided. All participants resided in regions where academic participation of this nature aligns with standard institutional practices. No personally identifiable information

was collected. The participants are informed and agree that the results of their validation will be used in this paper.

Validation Instructions. The human reviewers were provided with the following prompt and guidelines via an internal verification interface:

> *Task Instruction:* You are provided with an initial video frame, a textual task instruction, and six shuffled candidate frames. You are also given the "Ground-Truth Sequence Order" and the "GPT-5 Generated Reasoning Trace."
> *Verification Rules:*
> 1. Visual Check: Mentally simulate the physical transformation based on the task prompt. Does the provided ground-truth sequence represent a physically plausible and causally valid temporal evolution? (Yes/No)
> 2. Logic Check: Does the provided text rationale accurately explain the temporal progression without hallucinating non-existent objects or violating physical laws? (Yes/No)
> 3. Spatial Check: Do the '<LOCATE>' tags accurately correspond to the key visual regions undergoing dynamic changes? (Yes/No)
> If any answer is "No," flag the sample for removal or manual correction.

This strict protocol ensures that the evaluation benchmark is free from synthetic hallucinations and strictly adheres to verifiable physical laws.

Data Processing Prompt

You are an expert AI annotator specializing in physical reasoning and fine-grained spatio-temporal visual tracking. You are provided with the 'First frame', 6 candidate frames (labeled A through F) depicting a continuous physical event or transformation, AND the 'Ground Truth Chronological Sequence' (e.g., First frame - C - E - B - D - A - F).
Your task is to analyze the frames strictly following the provided ground truth order, identify the key changing objects or dynamic states driving the temporal progression, and generate dense, region-grounded reasoning that explains this exact sequence. You must strictly adhere to the following output structure and rules:
Instructions & Rules:

1. **1. Analyze the Provided Sequence:** Briefly explain the overall physical plausibility and causal logic governing the transformation from the first frame to the final state, based on the given sequence.
2. **2. Generate Verb-centric Semantic Cues:** You must track the localized dynamic changes across the candidate frames. When describing a key visual state or motion that justifies a frame's position in the sequence, you MUST enclose the descriptive text in `<LOCATE>...</LOCATE>` tags.
3. **3. The Action Rule:** The text inside the `<LOCATE>` tag MUST be a verb-centric phrase describing an action or a changing state (e.g., `<LOCATE>the red block being lifted into the air</LOCATE>` or `<LOCATE>the faces moving downwards and converging</LOCATE>`). Do NOT use static nouns (e.g., `<LOCATE>red block</LOCATE>` is invalid).
4. **4. Bounding Box Grounding:** For every `<LOCATE>` tag you generate, you must output the corresponding exact bounding box of that changing/moving object in the specified frame, using the format `[x_min, y_min, width, height]`.

Required Output Format:
Step 1: Temporal Logic of the Provided Sequence
[Provide the overarching causal logic explaining the given sequence]
Step 2: Grounded Visual Cues and Box Coordinates

- Following the sequence, Frame [Letter] shows `<LOCATE>`[verb-centric visual evidence]`</LOCATE>`. Box: `[x_min, y_min, width, height]`

- Next, Frame [Letter] captures `<LOCATE>`[verb-centric visual evidence]`</LOCATE>`. Box: `[x_min, y_min, width, height]`

... (Repeat for all key frames in the provided order)

Figure 9: **Data Processing Prompt.** The prompt template conditioning GPT-5 on ground truth sequences to generate verb-centric spatial-temporal reasoning chains and precise bounding boxes.