VLT: Vision-Language-Thinking (VLT) for Abstract Intent-Causal Planning (AICP) in Embodied AI via Human Demonstrations

Authors: Feng Lu, Feng Lu

Date: 2026-04-28T17:13:29+00:00

Abstract

Current end-to-end embodied models (VLAs) attempt to map perception directly to actions using a single network. However, due to their single optimization objective and sparse data distribution, these models degenerate into "trajectory replay machines" that rely on visual-action statistical shortcuts, lacking explicit causal reasoning capabilities and out-of-distribution generalization. To address this limitation, this paper proposes the Vision-Language-Thinking (VLT) paradigm, which extracts intents and physical causality from first-person human demonstrations and outputs Abstract Intent-Causal Planning (AICP). Instead of generating low-level trajectories, VLT produces sequential atomic-level abstract operations and subgoals; a supporting "cerebellum" execution layer is responsible for lightweight mapping and high-frequency closed-loop control. This architecture achieves the decoupling of "cognition and execution," providing a new interpretable, composable, and highly generalizable path for embodied intelligence.

Full Text

Preamble

Vision-Language-Thinking (VLT)

Abstract

Intent-Causal Planning (AICP) Embodied Human Demonstrations

1 (School of CSE , Beihang University, Beijing 100 191 , China)

Email: ORCID:

Abstract

Current end-to-end embodied models (VLAs) attempt perception directly actions using single network.

However, their single optimization objective sparse distribution, these models degenerate “trajectory replay machines” visual-action statistical shortcuts, lacking explicit causal reasoning capabilities out-of-distribution generalization. address limitation, paper proposes Vision-Language-Thinking (VLT) paradigm, which extracts intents physical causality first-person human demonstrations outputs

Abstract

Intent-Causal Planning (AICP). Instead generating low-level trajectories, produces sequential atomic-level

abstract

operations subgoals; supporting “cerebellum” execution layer responsible lightweight mapping high-frequency closed-loop control. architecture achieves decoupling “cognition execution,” providing interpretable, composable, highly generalizable embodied intelligence.

Keywords

Embodied Vision-Language-Thinking (VLT)

Abstract

Intent-Causal Planning (AICP) Human Demonstration;

1. Paradigm

Dilemma: “Shortcut Learning”hypothesis (Vision-Language-Action) “perception instructions directly actions.” However, practical training, several critical issues arise hinder development

cognitive capabilities:

First, single-objective optimization suppresses reasoning.

Behavioral cloning diffusion losses minimize trajectory errors without penalizing logical mistakes. result, models faster memorize spurious features background, lighting, fixed poses reduce losses learn physical causality phenomenon known Clever effect, where model appears perform correctly relies irrelevant rather genuine understanding.

Second, sparsity leads overfitting. Robot demonstrations cover extremely narrow "tubular manifold" state space, lacking negative samples boundary annotations.

Consequently, model learns successful snippets rather causal invariance, which essential generalizing unseen scenarios.

Third, black-box interface feedback. Discrete action tokens continuous trajectory outputs intermediate supervision, making impossible align semantic prior latent space action generation. leads paradoxical situation where model "fails brain (cognitive capabilities) struggles compensate non-existent cerebellum (fine-grained control)." Conclusion: inherent capabilities; instead, their architecture optimization objectives force bypass cognition directly trajectories.

Embodied intelligence requires clear division

labor between "cognition and control."

2. Cognitive

Reconstruction: Paradigm Atomic Planning (Vision-Language-Thinking) shifts optimization objective "fitting actions""understanding logic,"output interface being (Abstract Intent-Causal Planning). paradigm reformulates embodied models acquire process knowledge, detailed below:

Reconstruction Sources: Instead collecting robot trajectories, adopt first-person natural operation demonstrations human experts.

Equipped smart glasses wristbands, synchronously record movements, gestures, voice, movements, scene audio-visual data. approach preserves attention allocation, fault-tolerance adjustments, implicit experience real-world tasks i.e., information often traditional robot demonstration data.

Cognitive Inversion: Through multi-modal alignment, establish association graph between "environmental state changes operation effects human behaviors/attention." process strips implicit intents (e.g., anti-fall, precision preservation) physical constraint boundaries, extracting triplet "intent hierarchy behavioral logic environmental feedback." triplet forms foundation cognitive capabilities, enabling model, example specifically designed world model, grasp underlying logic tasks rather superficial trajectory patterns.

Output: output joint angles end-effector trajectories.

Instead, generates sequential specific subgoals atomic-level

abstract

operations (e.g., "reach toward handle," "lift vertically," "lower place gently") based aforementioned cognitive behavioral logic These units inherently carry ori-

entation, constraint boundaries (force/velocity/tolerance), temporal logic, completing “grounding

Abstract

physical intention” without mapping specific actuator space.

3. Execution

Closed-Loop: Cerebellum Architecture High-Frequency Physical Grounding demonstration trajectories.

cognitive iteration.

Optimization

Minimize trajectory error

Objective Output tokens (black Generalization Dependent coverage Mechanism Maximize intent consistency causal interpretability Atomic-level

abstract

operations constraint (gray Dependent causal invariance skill composition Since outputs already atomic units, cerebellum require secondary planning complex translation. primary responsibilities lightweight mapping, high-frequency execution, physical feedback, forming reliable execution closed-loop Minimalist Interface: cerebellum receives atomic instruction quickly generates control targets through built-in basic mapping rules (position/force control switching, reference target injection). eliminates long-range trajectory stitching, completely avoiding overfitting specific High-Frequency Closed-Loop:

Adopting classical control theory (impedance/admittance control, Model Predictive Control (MPC), Proportional-Integral-Derivative (PID) control), cerebellum operates frequency 100~1000Hz. integrates multi-source feedback force sensors, tactile sensors, encoders real-time compensate disturbances, maintain compliant contact, incorporate reflex-level emergency suppression mechanisms thereby ensuring stable execution dynamic environments.

Feedback Iteration: During local execution, cerebellum dynamically adjusts control parameters; after completion operation, compresses state

summary

(execution deviation/constraint violation/contact events) transmits

summary

verify whether “intent-causal chain”valid, deciding whether issue segment atomic instructions correct internal logic. forms closed-loop:

abstract

planning atomic mapping high-frequency execution state feedback

4. Paradigm

Advantages Interfaces Dimension (Action-Oriented) (Thinking-Oriented) Continuous trajectories discrete

Dimension (Action-Oriented) (Thinking-Oriented) Perception-action coupling, mutual

Cerebral cognition / cerebellar execution,

Control Division interference clear division responsibilities Post-hoc interception hard-coded High-frequency force feedback reflex- Safety Boundaries rules level emergency during execution Design: adopts structured protocol (JSON/Proto/Skill DSL), allowing cerebellum control algorithm pluggable mapping rules incrementally learned. bound specific models, sensors, controller types, reserving sufficient future advancements multi-modal fusion, self-supervised physical verification, reinforcement learning fine-tuning.

Conclusion

bottleneck embodied intelligence “larger models” “clearer division labor.” patches redefinition optimization objective embodied intelligence: model first learns output atomic-level intent-causal planning, cerebellum completes high-frequency, deterministic, force-feedback-enabled physical grounding, embodied system truly cross boundaries laboratory toward physical world. providing cognitive support, defining interface, cerebellum ensuring execution, drive embodied intelligence “demonstration replay” “verifiable, composable, evolvable cognitive agents.”

Note: Figure translations are in progress. See original paper for figures.