

QIAO-Bench: Predicting What Happens Next in Agent Workflows

From Trajectory Toward World Models for Reliable Long-Horizon Agents

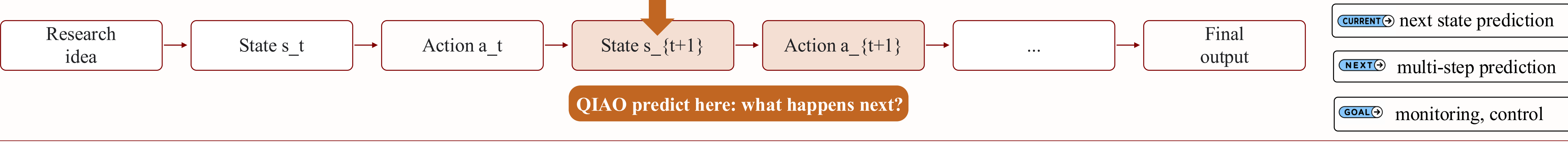
Qianqian He | Shane Gu | Haokun Liu | Chenhao Tan

University of Chicago CHAI Lab



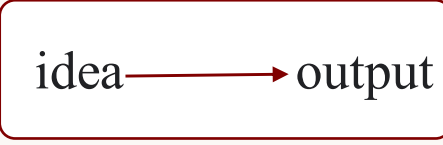
bellahe@uchicago.edu
chicagohai.github.io

From observing trajectories to predicting workflow dynamics



1. Problem & Hypothesis

Outcome-only evaluation hides how long-horizon agents reach the final output.



Hypothesis

Trajectory history supports prediction above random; action-conditioning can strengthen learned predictive representations.

2. QIAO-Bench

QIAO = Quantifying Information, Actions, and Outcomes for Efficient Hypothesis Verification

Information

state + recent events

Actions

tool / command / query

Outcomes

future phase + action

1,126

decision points

9

autonomous runs

40

phase-action classes

107

semantic classes

Quality tiers: 502 gold | 375 silver | 249 bronze

Group split by complete runs: 5 train | 2 validation | 2 test

3. Experimental Design

Random retrieval is the baseline; generalization is measured on held-out runs.

Compare to random

Test transfer

Inspect signal

Identify model

Inputs

state only, action only, state + action

Predictors

random, TF-IDF, JEPA-inspired, ridge

Targets

phase-action, exact record, ranking

Metrics

recall@1, exact recall@10, MRR

4. Scope And Limitations

QIAO-Bench is the first milestone of our benchmark scope, current limits:

- 9 independent runs
- no causal action effects yet
- automatic label may be noisy
- no downstream intervention utility yet
- one-step prediction only

Key Claim

Trajectories contain predictive structure. Need more tests for control and intervention.

5. Key Findings

Predictable

future workflow structure transfers

Action conditioned

actions improve learned prediction

Need control

causal utility & intervention tests

Key Takeaway

Next-state prediction is a plausible first layer for future monitoring, recovery systems, and action-conditioned world models.

6. Framework: Trajectories → Representation → Prediction

Observed

Infrastructure

agent runtime, tools, files, memory, model, environment

Observed

Trajectory extraction

$\tau = \{(s_t, a_t, o_{t+1})\}$ decision points

Represent

Canonical features

stage, phase, recent events; tool, command, query, file

Model

Predictive representation

TF-IDF/SVD + latent encoders; JEPA-inspired prediction

Test

Evaluation

held-out run retrieval: R@1, Exact R@10, MRR

Future

Application

monitoring, recovery, intervention, resource-aware control

Key Contributions

QIAO-Bench converts traces into held-out next-state prediction benchmark. It links information, action, and outcome at each decision point.

7. JEPA-Inspired Predictor

s_t

State

a_t

Action

Gated Fusion

Future Latent

Toward World Models

One-step prediction now; multi-step rollout for the next.

➤ State/action encoders: $z_s = E_s(s_t)$, $z_a = E_a(a_t)$

➤ Gated fusion: $\hat{z}_{t+1} = P(z_{sa})$

➤ Future latent prediction: $z_{sa} = z_a + \sigma(\alpha) \cdot g([z_s; z_a]) \odot z_s$

➤ Training: $L = L_{\cos} + 0.1 L_{NCE} + 0.5 L_{\text{phase}} + 0.5 L_{\text{action}}$

8. Result & Evidence

Hypothesis is supported: trajectories are predictive, but no representation dominates.

Action-only:
JEPA helps most

0.214 ± 0.075

JEPA R@1



0.108 ± 0.056 /
 0.077 ± 0.034

Ridge / TF-IDF R@1

State + Action:
JEPA helps most

0.294 ± 0.022

JEPA R@1



0.247 ± 0.018 /
 0.210 ± 0.059

TF-IDF / Ridge R@1

State-only: workflow structure is strong

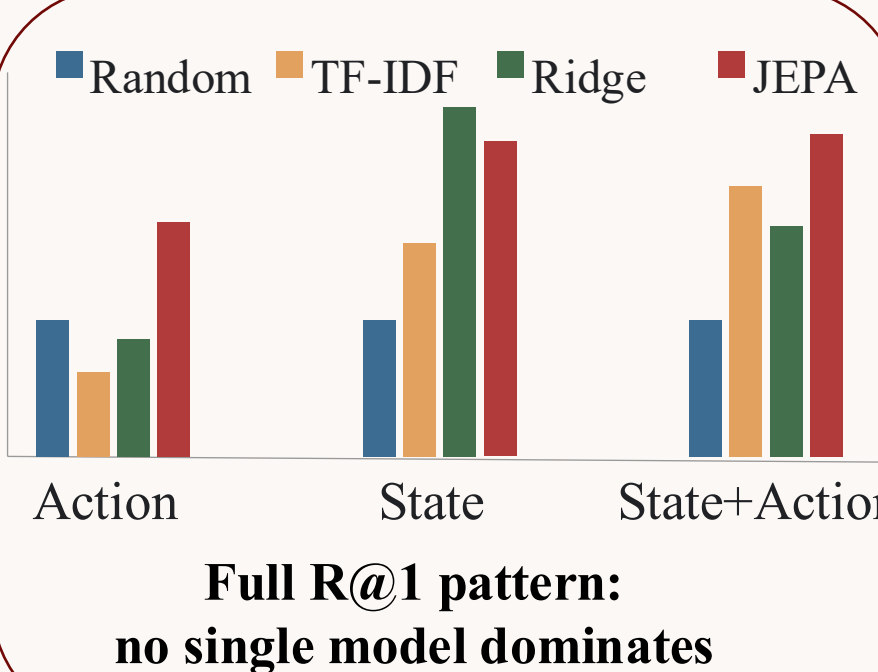
0.319 ± 0.063

Ridge R@1



0.287 ± 0.016 /
 0.195 ± 0.097

JEPA / TF-IDF R@1



Input	Best	R@1	Content
Action only	JEPA	0.214	latent action signal
State only	Ridge	0.319	workflow structure
State+Action	JEPA	0.294	combined context

Best by input: no single model dominates

Roadmap: From Next-Stage Prediction To Reliable Long-Horizon Agent Systems

QIAO-Bench
next-state prediction

Temporal modeling
multi-step trajectory prediction

Predictive monitoring
failure / anomaly /wrong decision forecasting

Reliable agent systems
recovery/intervention/
resource-aware execution

NOW

GOAL

Decision modeling
counterfactual action ranking

QIAO-Bench is the first empirical layer.

The long-term objective is to move from prediction to monitoring, recovery, intervention, and resource-aware execution for long-horizon agent workflows.