ActWorld: From Explorable to Interactive World Model via Action-Aware Memory

Navigation-only world models **forget actions**. ActWorld organizes memory by **interaction events, not recency**.

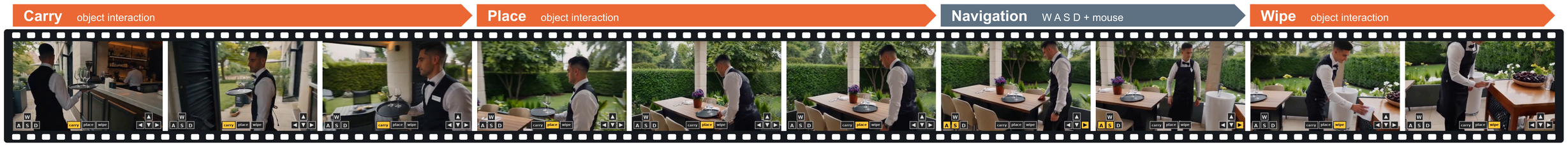
Zhexiao Xiong[1,2,*], Yizhi Song[2], Hao Kang[2], Qing Yan[2], Liming Jiang[2], Jenson Yang[2], Zhoujie Fu[2], Stathi Fotiadis[2], Zichuan Liu[2], Bo Liu[2], Yiding Yang[2], Xin Lu[2], Nathan Jacobs[1]

[1]Washington University in St. Louis [2]Intelligent Creation, ByteDance

*Work done during internship at ByteDance

Project page | Code | Personal webpage

On the job market: Zhexiao Xiong is seeking full-time Research Scientist / Applied Scientist / Member of Technical Staff roles.



1 Explorable ≠ interactive

- Interactive world models today only **navigate**.
- **Object interaction** is absent, game-only, or offline.
- Two bottlenecks: **data** and recency-based **memory**.

Action-forgetting: recency vs. importance

contact / manipulating chunk | navigation chunk | event / object token

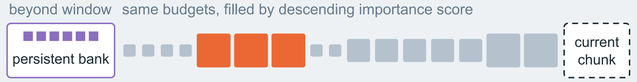
square size = tokens per chunk

Helios (base): age picks the channel



→ *Contact chunks age into coarse: the lamp you switched on is forgotten.*

ActWorld: interaction-importance score picks the channel



→ *Old but causal contact chunks stay fine; idle navigation chunks are demoted.*

2 Data & annotation

100K	40	6	400 h
interaction videos	action categories	interaction phases	real navigation (Sekai)

Chain-of-thought per-chunk annotation

33-frame chunk (≈1.4 s) → 5 key frames → VLM chain-of-thought: what changed? contact? which phase? → **interaction flag phase label dense caption**

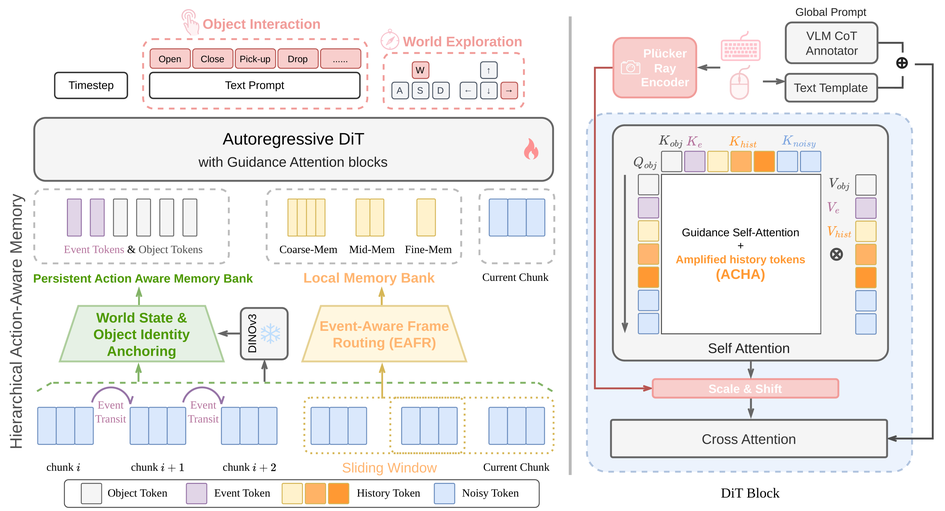
Six-phase taxonomy tells the memory what matters

approaching | reaching | **contact** | **manipulating** | completing | post-action

contact / manipulating frames decide later object states.

3 Method: hierarchical action-aware memory

Chunk-autoregressive DiT on Helios-14B · Plücker-ray camera control · distilled to a 3-step real-time generator



EAFR Event-Aware Frame Re-assignment

Fill buckets by **importance**, not age.

$$w_k = \lambda_\varphi \, \varphi\left(y_k^{\text{int}}, y_k^{\text{ph}}\right) + \lambda_r \, e^{-(t-k)/\tau}$$

φ: phase prior peaking at **contact / manipulating**; a recency term remains. RoPE time index is kept.

Ablation: action success 52.9 → 53.8 %

ACHA Action-Conditioned History Amplification

Amplify history keys for the **current action**.

$$K'_{\text{hist}} = \text{softplus}\left(s + W_2\,\sigma(W_1 \mathbf{e}_a)\right) \odot K_{\text{hist}}$$

Per-head amplifier in self-attention; zero-init, so it equals the backbone at step 0.

Ablation: instruction following 2.35 → 2.41

Persistent action-aware memory bank survives the window's eviction horizon

Event tokens — what happened, and when

$$\mathbf{e}_k^{\text{evt}} = E_\xi[\xi_k] + E_{\text{ph}}[y_k^{\text{ph}}] + E_a[a] + \text{AttnPool}(\mathbf{h}_k)$$

Written at phase transitions:
ENTER-MANIP · ENTER-COMPLETE · RELEASE

Object tokens — what the touched object looks like

$$\mathbf{e}_{k,j}^{\text{obj}} = W_v\left(\mathbf{f}_{k,j}^{\text{dino}}\right) + E_{\text{ph}}[y_k^{\text{ph}}] + E_a[a]$$

Top-3 salient DINOv3 patches per interaction chunk; view-stable across long navigation gaps.

16 tokens, FIFO, interaction tokens pinned. **Ablation:** success 54.0 → 57.8 %, subject consistency 0.844 → 0.871

4 I-Bench: our new benchmark

Existing benchmarks test roaming **or** fixed-camera interaction. I-Bench chains **both in one long clip**.

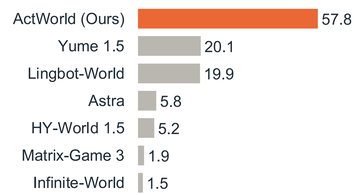
300	3 + 2–3	10 × 33	50 / 50
prompts in 30 sequences	actions + camera moves per prompt	chunks × frames per clip	first- / third-person views

3 axes: **VLM-Action-Judge** does the action happen? · **Key-Mouse-Following** does the camera obey? · **VBench** visual quality

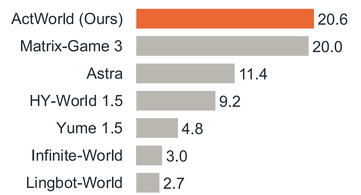
Results vs. 6 baselines (official checkpoints)

57.8%	4.05 / 5	7 / 10
action success: 2.9× the best baseline (20.1%)	user-study action following (best baseline 2.68)	VBench dimensions where ActWorld is best

Action success VLM-AJ (%)



Key + mouse following Acc. full (%)



Pour → Stir: baselines lose the objects; ActWorld keeps them

