When Does JEPA Planning Fail? A Precise Boundary Condition from 2D Billiards

Santosh Jaiswal
Independent Researcher • github.com/hellojais • WM@Booth Chicago 2026, Submission #2

Why Billiards?

- **Sharp physics boundary:** collision compresses a velocity non-linearity into a *single frame* – the hardest case for JEPA's smooth prediction objective.
- **Low visual complexity:** two balls on a plain background give SIGReg little signal to organise embeddings by task-relevant structure.
- **Controlled benchmark:** fully observable, deterministic physics, ground-truth state at every step. Unlike Push-T (quasi-static, visually rich), billiards requires microscopic timing at impact.
- **Parallels Two-Room:** LeWM flags low intrinsic dimensionality as a planning boundary; we characterise the mechanism precisely.

Dataset & Architecture

Dataset: 4,000 episodes, 971,321 frames.
Each timestep: 96×96×3 RGB pixels, 2D action impulse, 10-dim state (cue/target x, y, v_x, v_y; pocket x, y).

Expert: ghost-ball $G = T - \hat{u}_{PT} \cdot 2R$ with Gaussian noise; 56% episode success rate; failed episodes retained.

Shared encoder (all experiments): ViT-Tiny, patch 8, 96×96, hidden 192, 12L, 3H.
Projector: $192 \to 2048 \to 32$-dim.
CEM planner: H=10, N=1000 samples, 20 iterations.

Probe R^2: What Is Encoded?

MLP ($32 \to 64 \to 10$) on frozen 32-dim post-projector embeddings.

Dimension	Baseline	FrameStack	AuxLoss
cue x / y	0.854 / –	0.209 / 0.366	**0.998 / 0.997**
cue v_x / v_y	– / –	0.008 / 0.023	**0.971 / 0.968**
target x / y	0.988 / 0.976	0.297 / 0.596	**0.999 / 0.998**
target v_x / v_y	0.337 / 0.371	0.457 / 0.064	**0.927 / 0.924**
pocket x / y	0.981 / –	0.244 / 0.177	**0.998 / 0.999**
pos mean	0.983	0.446	**0.999**
vel mean	0.296	0.138	**0.947**

Baseline encodes position (R^2=0.988), barely velocity (R^2=0.296). AuxLoss recovers all ten dimensions.

Two Sequential Bottlenecks for JEPA Planning

Bottleneck 1 – Encoder ⇒ **Fix – AuxLoss**

Baseline: vel R^2=0.296 (barely encoded).
FrameStack evicts position:
pos R^2: 0.983 → 0.446 (eviction)

pos R^2=0.999, vel R^2=0.947
3.2× lower prediction loss

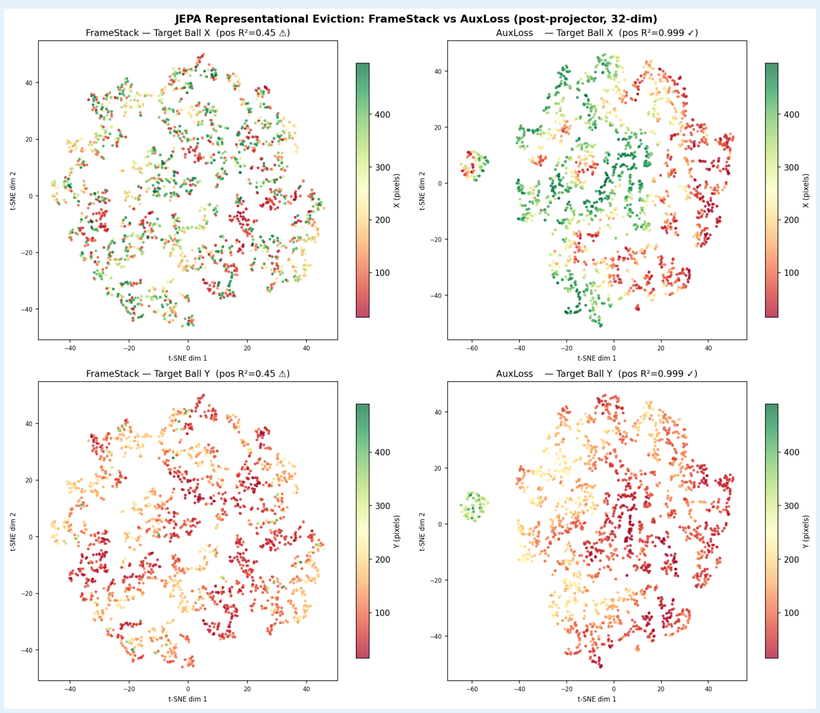
⇓ **Encoder fixed – but planning still fails**

Bottleneck 2 – Predictor ⇒ **Open Problem**

Rollout accuracy collapses at collision timesteps

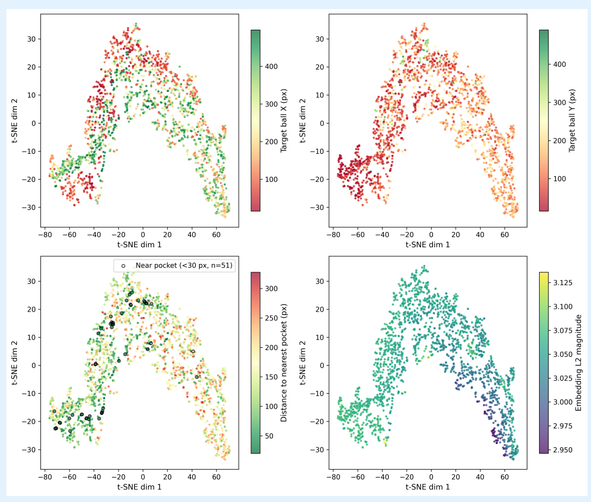
Pocket-proximity CEM: 28% progress
State-based hybrid: **5/5 success**

Representational Eviction vs. AuxLoss Recovery



t-SNE of 32-dim embeddings coloured by target ball x (top) and y (bottom). **Left (FrameStack):** explicit optical flow causes SIGReg's fixed-capacity regularisation to trade absolute position (R^2: 0.983 → 0.446) for velocity – **representational eviction**. Both quantities are needed for planning; neither alone suffices. **Right (AuxLoss):** auxiliary state head restores all structure (R^2=0.999) and reduces prediction loss 3.2×.

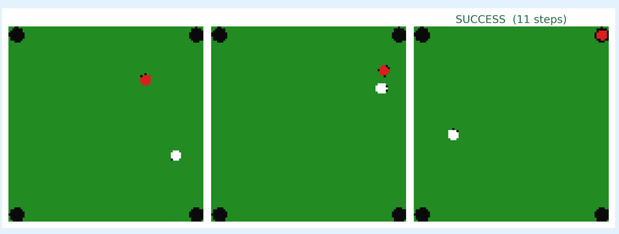
Baseline Latent Space: Why CEM Has No Gradient



Experiment Summary

Exp	Key finding	Outcome
1	Emb CEM 192-dim: uniform shell (CoV 0.081)	Fail
2	Emb CEM 32-dim: 8.5× better pred; cost still flat	Fail
3	State-based CEM: pots in 9–13 steps	**5/5**
4	Probe CEM: embs diverge at collision	Fail
5	Mamba: vel R^2 0.296→0.297; encoder is bottleneck	Fail
6	FrameStack: pos R^2 0.983→0.446 (eviction)	Fail
7–8	AuxLoss: pos R^2=0.999, vel R^2=0.947	Enc. fixed
9	Emb CEM + fixed encoder: zero progress	Fail
10	Pocket CEM + fixed encoder: 28% progress	Fail
11	Pocket CEM + FrameStack: flat, zero progress	Fail

Cross-Episode Planning: State-Based Hybrid



Combo B: start from episode 100, goal from episode 3,000 – a combination *never seen during training*. State-based CEM pots the ball in 13 steps (5/5 seeds). The model learned generalizable billiards physics; the failure is predictor rollout accuracy at collision timesteps, not the underlying representation.

Conclusion

- Position encoded at R^2=0.988 from raw pixels with no explicit supervision – JEPA representations are rich.
- **Bottleneck 1 (encoder):** SIGReg's isotropic regularisation suppresses velocity; FrameStack triggers representational eviction. AuxLoss resolves both.
- **Bottleneck 2 (predictor):** With pos R^2=0.999 and vel R^2=0.947, predictor error at collision timesteps still forecloses imagination planning. Open problem.
- **Diagnostic framework:** (1) Probe encoder R^2. (2) Test pocket-proximity CEM. Both must pass for pure JEPA planning to succeed in high-precision physics.

Resources

Dataset `hf.co/hellojais/billiards-worldmodel`
Model `hf.co/hellojais/lewm-billiards`
Code `github.com/hellojais/le-wm`
Findings `.../le-wm/blob/main/FINDINGS.md`

LeWM: Maes, Leroux, Gidel, Berseth (2025), arXiv:2603.19312