ICML
International Conference
On Machine Learning

Position:
Profiling Game Worlds by Transition Complexity

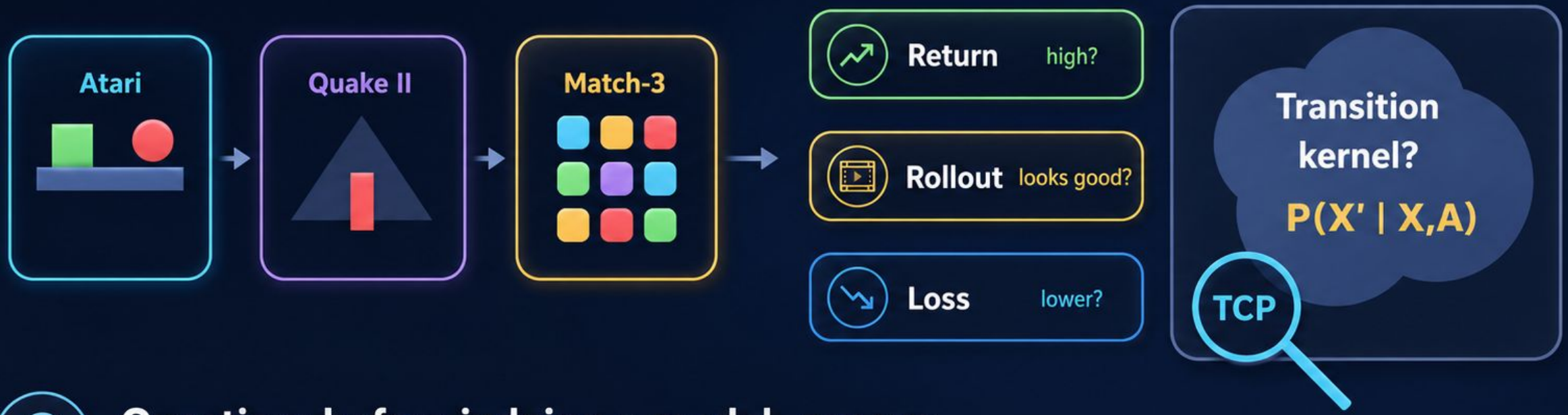
Lele Cao

King

ACTIVISION BLIZZARD

The hidden denominator

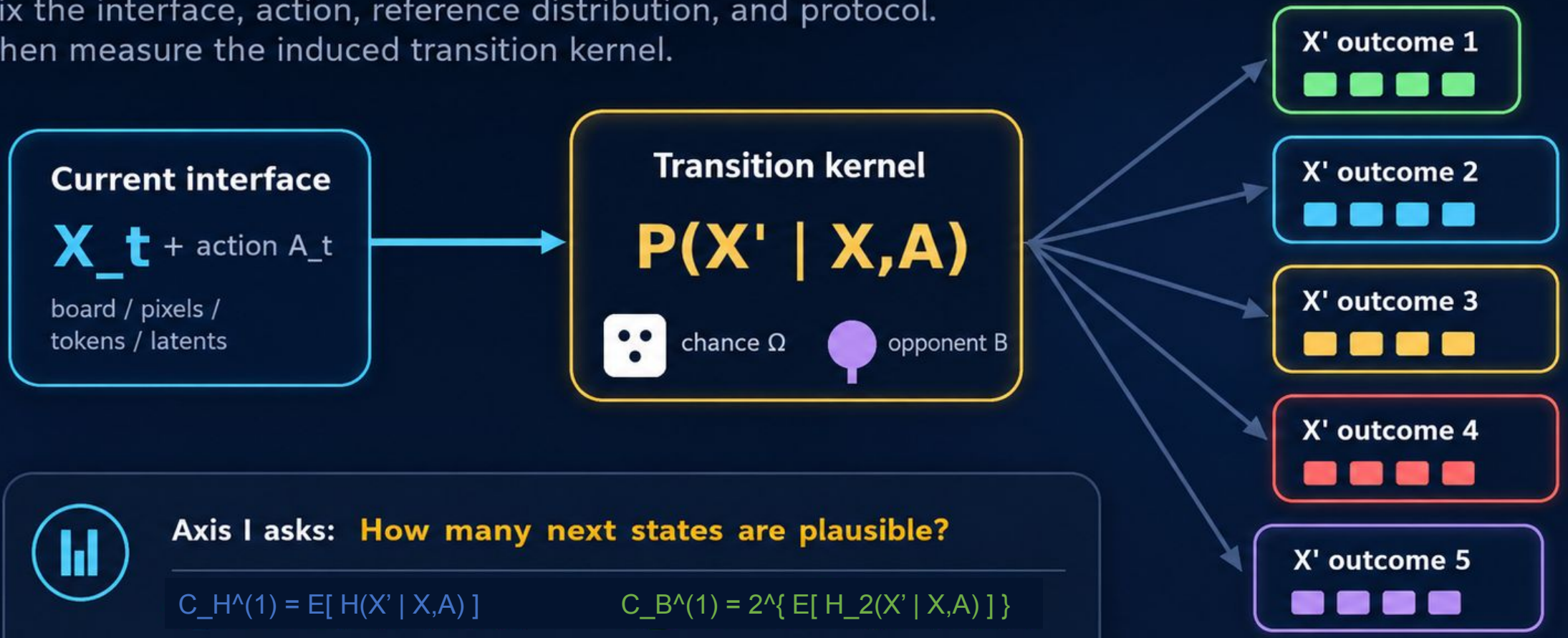
World models are compared by return, visual quality, or loss —
but those metrics rarely say how hard the transition problem was.



Question: before judging a model, can we measure the world it had to predict?

TCP profiles the next-step lottery

Fix the interface, action, reference distribution, and protocol.
Then measure the induced transition kernel.



Axis I asks: **How many next states are plausible?**

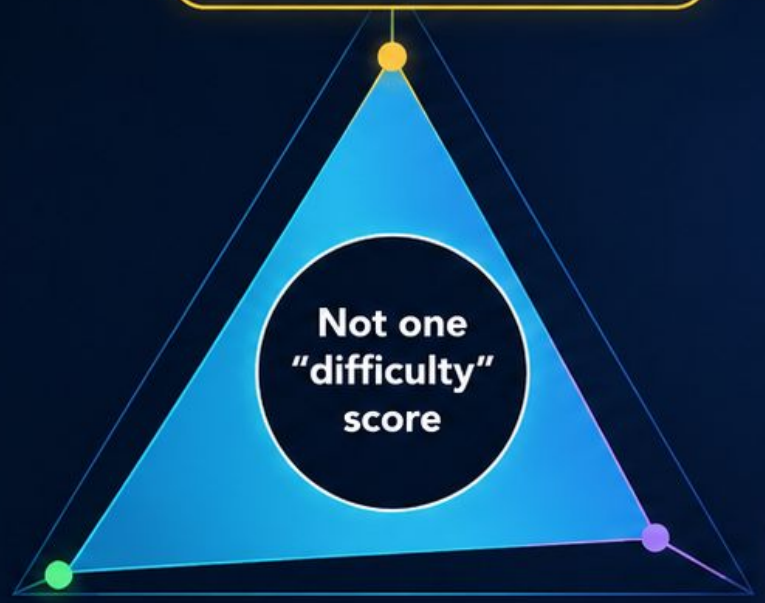
$C_H^{(1)} = E[\, H(X' \mid X,A) \,]$ (entropy-based)

$C_B^{(1)} = 2^{\{ E[\, H_2(X' \mid X,A) \,] \}}$ (collision-entropy based effective branching)

One scalar is too blunt

TCP is a profile: branching, interaction, and span are different failure modes for world models.

I Branching
How many next states?



III Span
How much history or neighborhood?

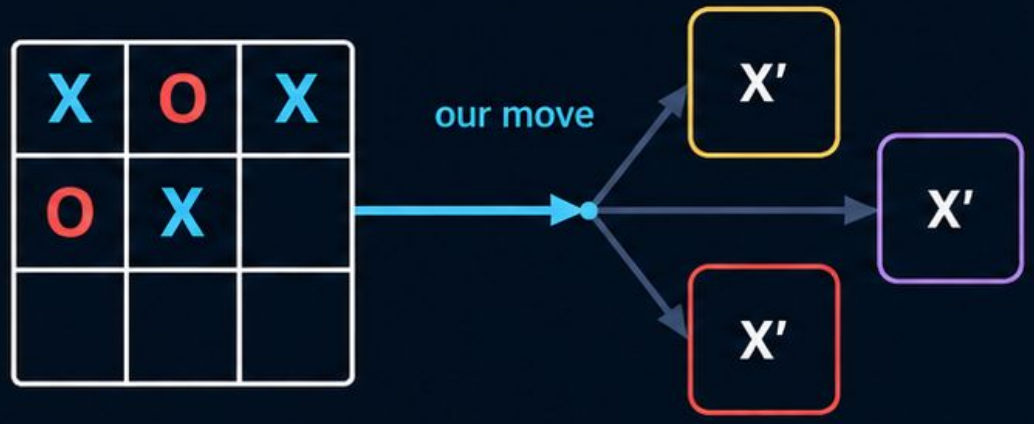
II Interaction
Who else can change it?

Narrow claim: profile the induced next-interface prediction problem — not rule-description complexity, solver difficulty, or return.

Two tiny examples make the point

Once X, d, and the step protocol are declared, TCP numbers become interpretable.

Tic-tac-toe: opponent branching



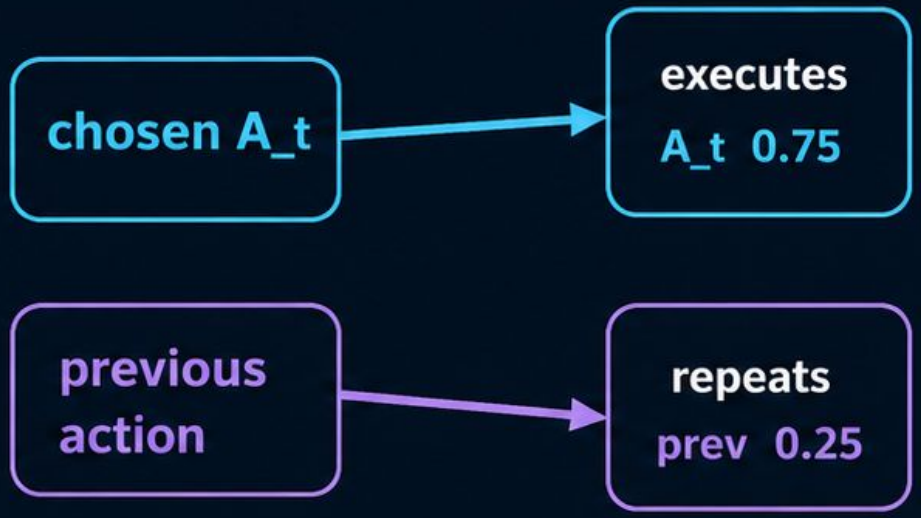
Random-play reference distribution

$C_H^{(1)} \approx 1.906$ bits

$C_B^{(1)} \approx 3.75$ effective branches

Here, branching comes from opponent replies.

Sticky action: protocol branching



If outcomes differ:

$h_2(0.25) \approx 0.811$ bits

The recipe is deliberately boring

TCP should be reproducible benchmark metadata, not a new hidden tuning axis.

TCP card

TCP card

Interface X_t	board / pixels / tokens / latents
Reference d	d_data and d_probe
Protocol	wrappers, resets, stickiness
Budget	sample count, M, probes, seeds
Axes	I branching II interaction III span

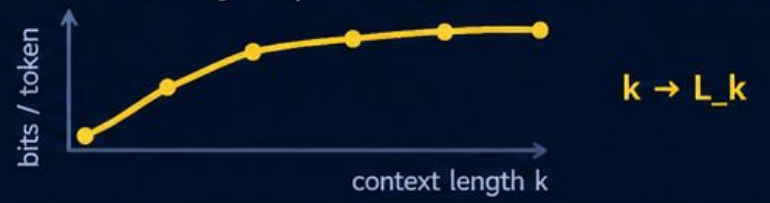
Simulator-backed

Reset to fixed (x,a), resample randomness/opponents, estimate collision entropy H_2.

Lite-v1: 5,000 pairs, M=8 | Std-v1: 20,000 pairs, M=32

Log-only

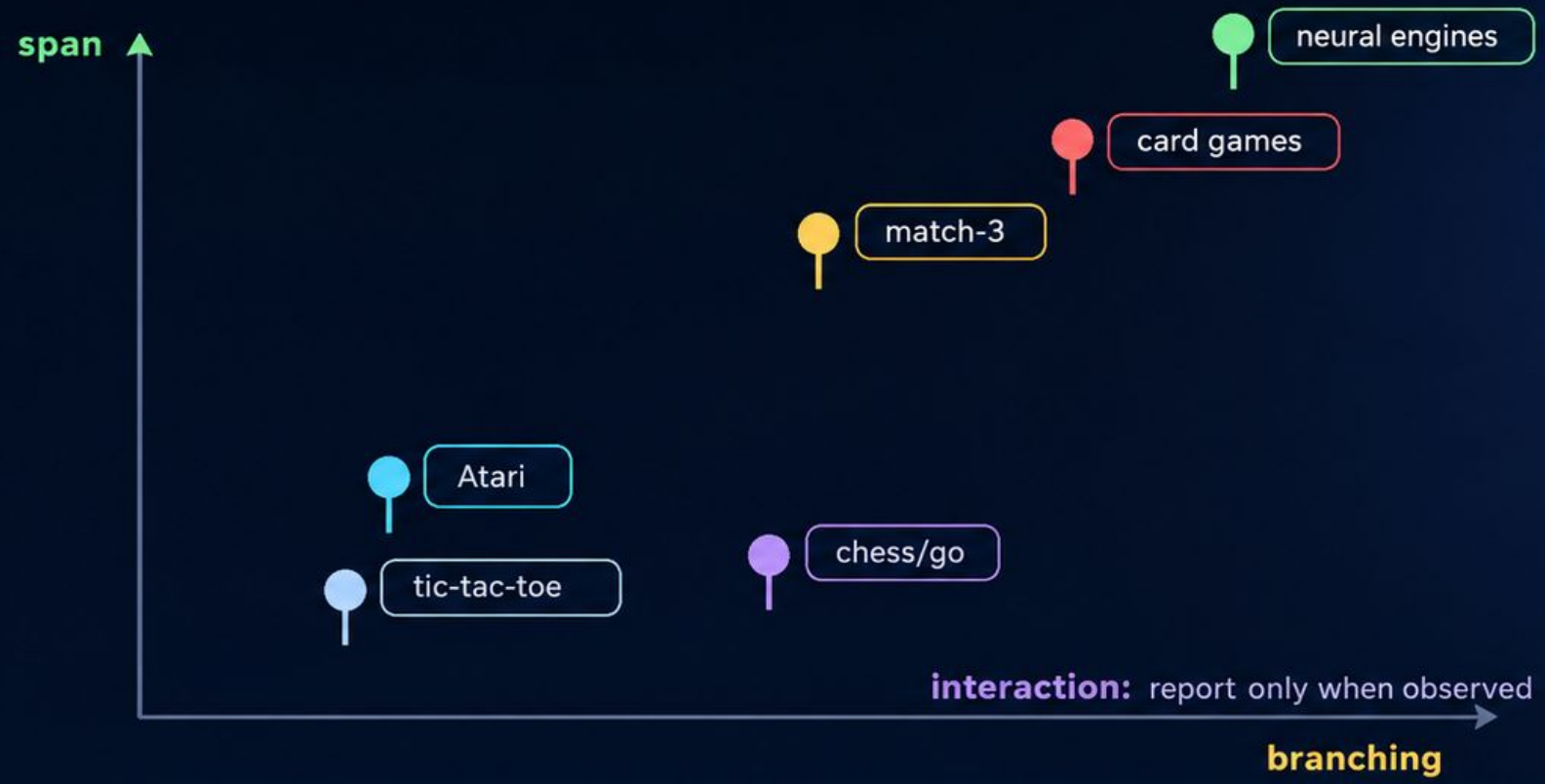
Train fixed TCP-Ref probes, report bits/token and memory-depth curves across context lengths.



GRU-v1 + TX-v1, 3 seeds, CIs, capacity sanity check

Make TCP benchmark metadata

Then return, rollout quality, and prediction loss get the context they were missing.



Who uses the card?

Benchmark maintainers
publish TCP-Std cards

Model papers
cite or recompute if kernel changes

Suite designers
span the TCP axes

Measure the world, then compare the model.