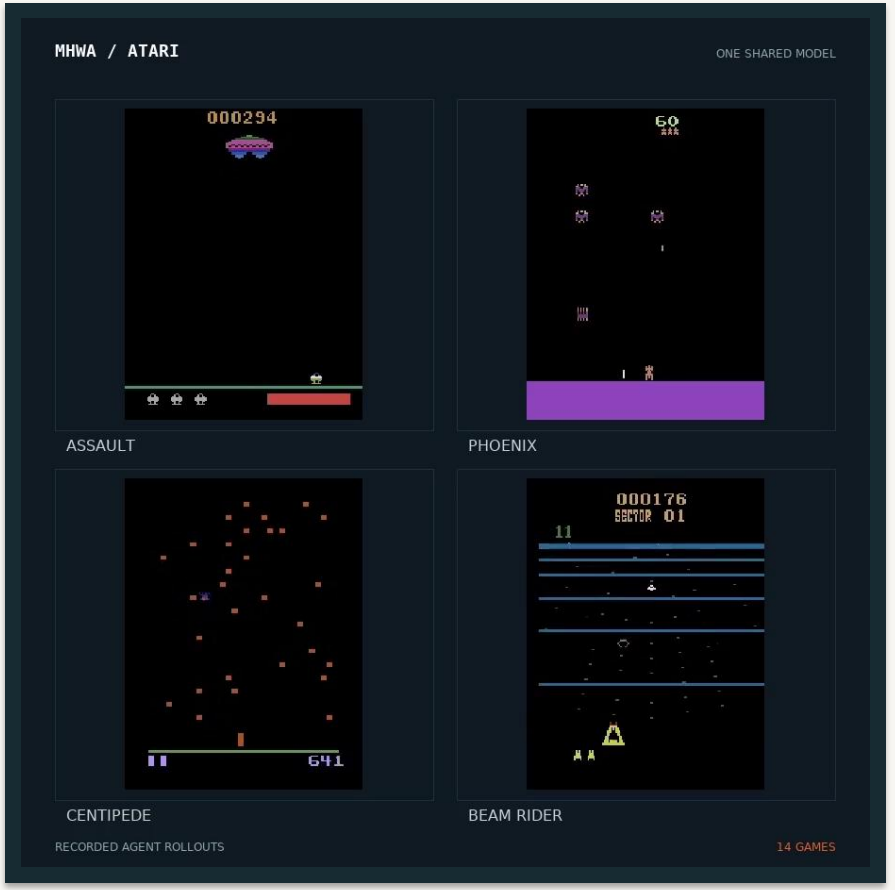
MHWA.

Multi-timescale Hierarchical World-Action Model

Learn the world.
Act with structure.

Pengcheng Pan, Guoqing Ma, Yuhan Zhang,
Yang Chen, Yichen Liu, Ziheng Li, Shan Yu

Institute of Automation, Chinese Academy of Sciences



ONE MODEL / 14 ATARI GAMES

Small prediction errors compound.

Can better structure reduce the need for greater scale?

REAL CONTEXT

Observed history

→

PREDICTION

Becomes the next input

→

LONGER ROLLOUT

Errors can accumulate

CONCEPTUAL INTERFERENCE

Unstable high-level meaning

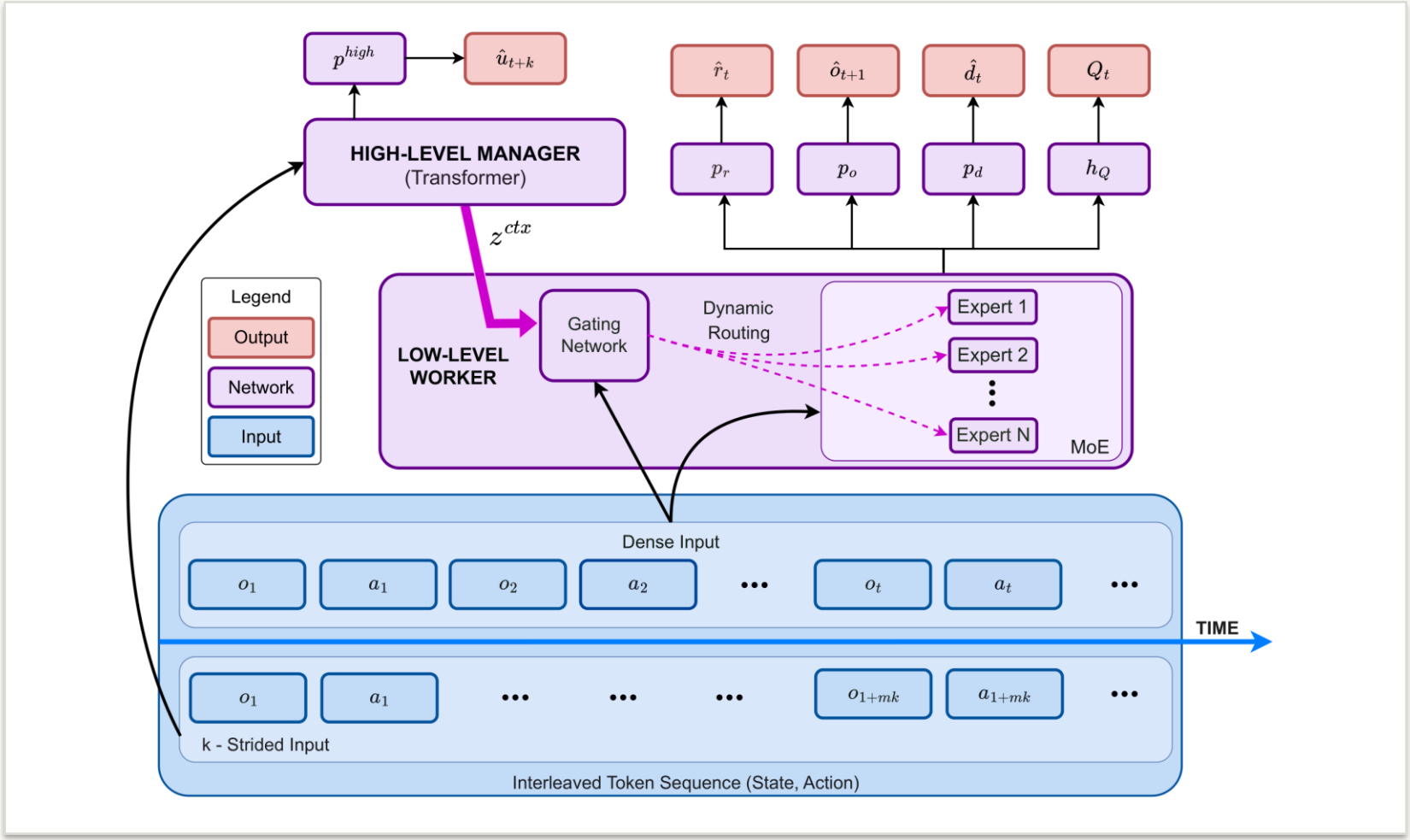
Low-level learning can disturb the context representation.

STRUCTURAL INTERFERENCE

One network, diverse dynamics

Shared low-level parameters must fit heterogeneous behaviors.

Two levels. One world-action model.



01 / MANAGER

Infer slow context

Sample input embeddings every k = 4 tokens.

02 / PREDICTION HEAD

Predict the next sample

The target token can be an observation or an action.

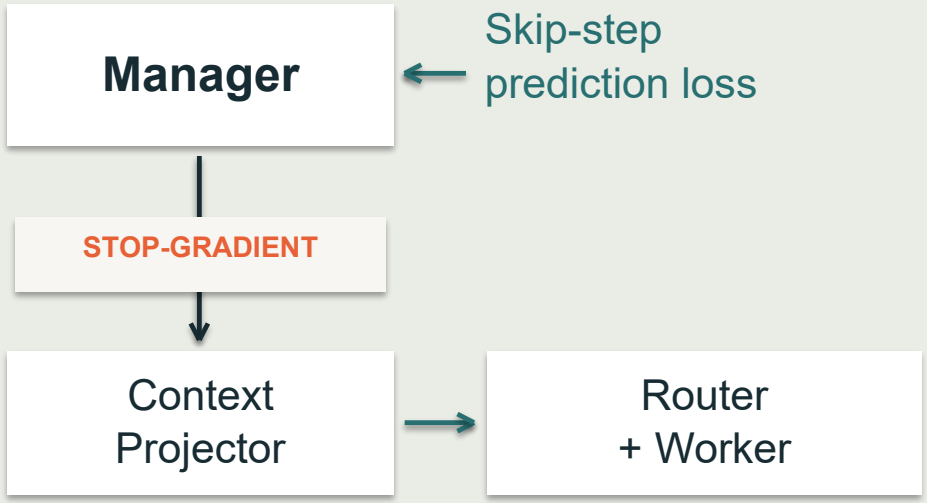
03 / WORKER GATE

Select and weight experts

Combine context + current token. Activate 2 of 4 experts.

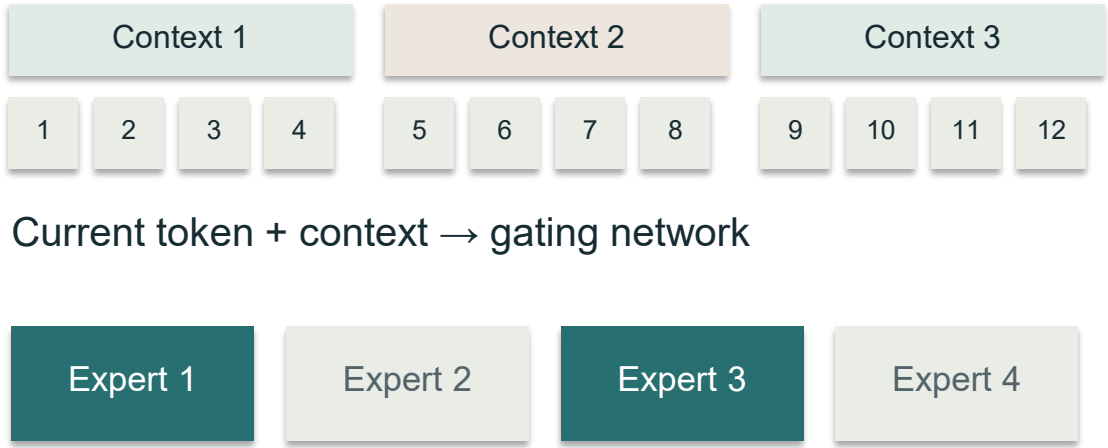
Separate learning. Coordinate computation.

ISOLATE THE MANAGER'S LEARNING



Projector, router and Worker remain trainable.

COARSE CONTEXT, TOKEN-LEVEL ROUTING



2 selected experts · normalized weights

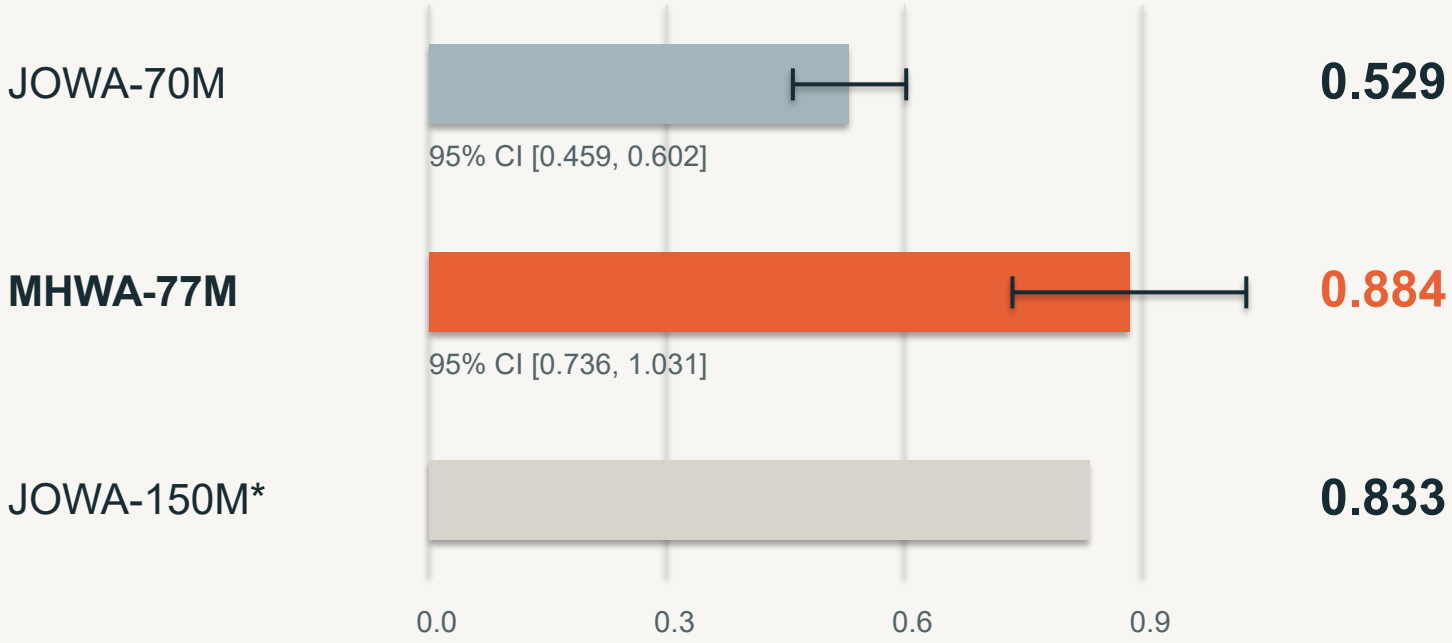
KL balance on batch-average routing probabilities

Illustrative token positions and expert selection. Default context refresh: r = k = 4.

Strong performance. Smaller training footprint.

14 Atari games · 10% offline data · 3 training seeds in the matched comparison

IQM HUMAN-NORMALIZED SCORE ↑



TRAINING GFLOPs / SEQUENCE

JOWA-150M	165.524
JOWA-70M	84.238
MHWA-77M	48.199

70.88%

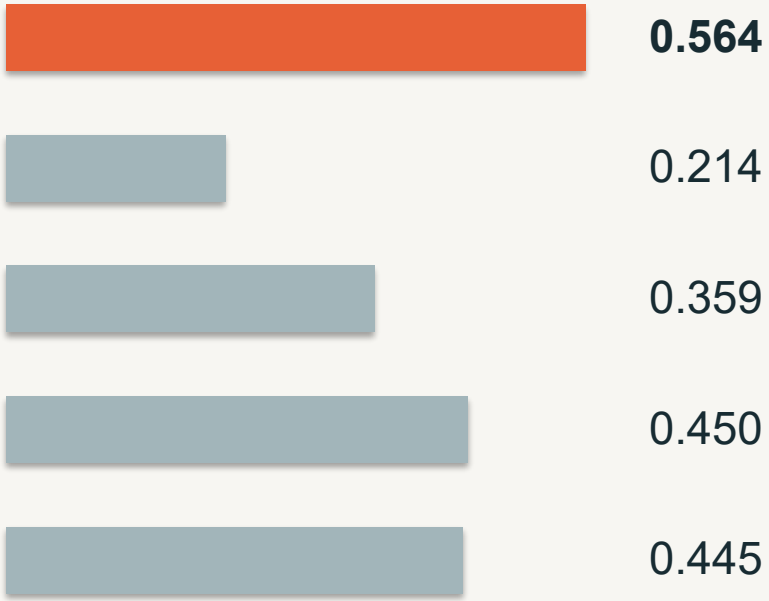
fewer FLOPs vs. JOWA-150M

IQM: RLiable seed-by-game aggregation. * Published reference, separate from the matched comparison.
FLOPs: forward + backward, identical inputs (8 frames / 296 tokens). Sources: CR Tables 2 and 6.

The components work together.

Controlled 44M-parameter model · 6-game subset · IQM HNS

ARCHITECTURAL CHOICES



HIGH-LEVEL SUPERVISION × GRADIENT ISOLATION

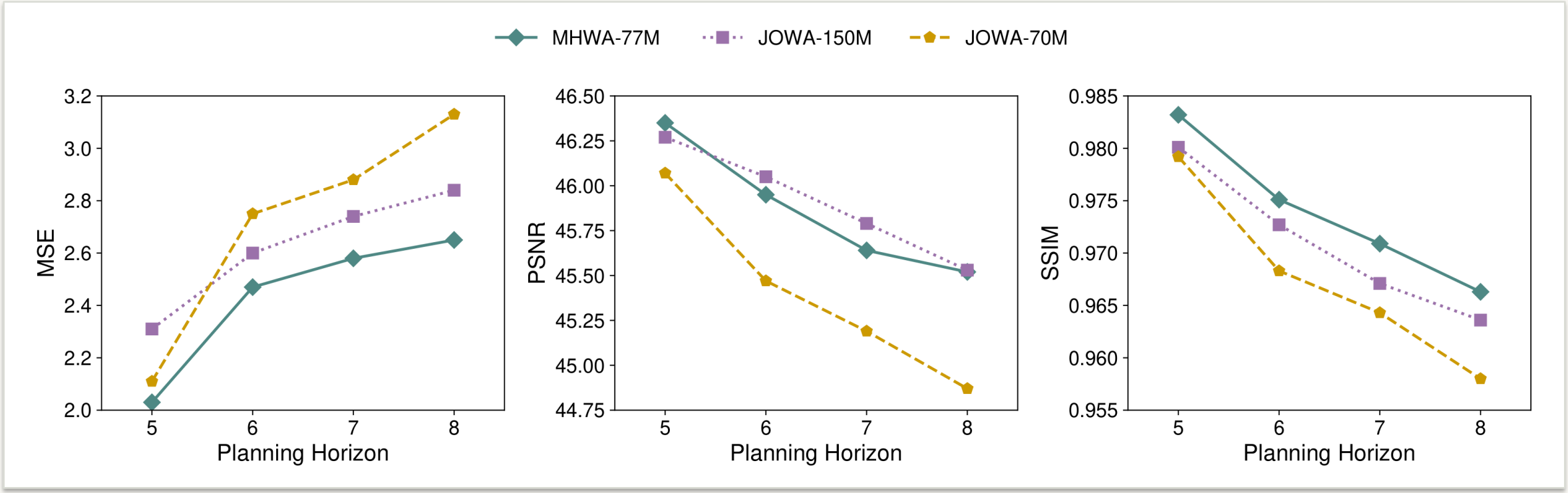
	SG on	SG off
High loss on	**0.564**	0.533
High loss off	**0.503**	0.387

Stop-gradient helps with and without skip-step supervision.

SG = stop-gradient. All four factorial settings keep the Context Projector and Worker trainable.
The 44M / 6-game ablation setting is separate from the main 77M / 14-game comparison. CR Tables 3–4.

Better predictions across rollout steps.

4 real context frames → open-loop predictions at steps 5–8 · recorded actions fixed



HELD-OUT ADAPTATION

0.643 ± 0.007 IQM DNS on 5 unseen Atari games

Reconstruction: final CR Figure 2. Held-out fine-tuning follows JOWA; JOWA-150M reference: 0.647.
Adaptation results: Appendix E, Table 9. DNS = DQN-normalized score.

Structure is another axis of scale.

Slow context.

Reactive experts.

Less training computation.

Adaptive timescales and broader control domains remain future work.

preplic.github.io/MHWA-project-page/

0.884

IQM HNS · seed-by-game

77M

parameters

70.88%

fewer training FLOPs vs. JOWA-150M

arXiv: 2610.14727