A Methods

A.1 Open source attention module

We have open-sourced a Jax/Haiku implementation of our Hierarchical Attention Module, which can be found at: `https://github.com/deepmind/deepmind-research/tree/master/hierarchical_transformer_memory/hierarchical_attention`

We have also released our Ballet and Rapid Word Learning environments, and include links to those and the released versions of other environments we used below.

A.2 Agent architecture and training

Table 1: Hyperparameters used in main experiments. Where only one value is listed across multiple columns, it applies to all. Where hyperparameters were swept, parameters used for HCAM and TrXL are reported separately.

	Ballet	**Objects**	**Words**	**Image**	**Streets**	**Associative**
All activation fns	ReLU					
State dimension	512					
Memory dimension	512					
Memory layers	4					
Memory num. heads	8					
HCAM chunk size	32	16	16	16	4	2
HCAM chunk overlap	-				1	-
HCAM top-k	8	8	16	16	32	32
HCAM total num. chunks	always greater than max episode length // chunk size					
Local Attention Window	64	128	128	128	64	NA
TrXL extra length	256	512	256	NA	200 (full)	NA
Visual encoder	CNN	ResNet				
Vis. enc. channels	(16, 32, 32)					
Vis. enc. filt. size	(9, 3, 3)	(3, 3, 3)				
Vis. enc. filt. stride	(9, 1, 1)	(2, 2, 2)				
Vis. enc. num. blocks	NA	(2, 2, 2)				
Language encoder	1-layer LSTM			NA		
Lang. enc. dimension	256			NA		
Word embed. dimension	32			NA		
Policy & value nets	MLP with 1 hidden layer with 512 units.					
Reconstruction decoders	Architectural transposes of the encoders, with independent weights.					
Recon. loss weight (HCAM)	1.		0.3	1.	0.3	NA
Recon. loss weight TrXL	1.			NA	1.	NA
V-trace loss weight (HCAM)	0.1		0.3	0.1	1.	NA
V-trace loss weight (TrXL)	0.3	0.1		NA	1.	NA
V-trace baseline weight (HCAM)	1.		0.3	1.	0.3	NA
V-trace baseline weight (TrXL)	1.	0.3	1.	NA	1.	NA
Entropy weight	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$				NA
Batch size	32					
Training unroll length	64	128	128	128	64	NA
Optimizer	Adam [27]					
LR	$2 \cdot 10^{-4}$				$4 \cdot 10^{-4}$	$1 \cdot 10^{-4}$