
SDMTR: A Brain-inspired Transformer for Relation Inference

Xiangyu Zeng

University of
Electronic Science and Technology of China

Jie Lin¹

University of
Electronic Science and Technology of China

Piao Hu

University of
Electronic Science and Technology of China

Zhihao Li

University of
Electronic Science and Technology of China

Tianxi Huang

Department of Fundamental Courses, Chengdu Textile College

Abstract

Deep learning has seen a movement towards the concepts of modularity, module coordination and sparse interactions to fit the working principles of biological systems. Inspired by Global Workspace Theory and long-term memory system in human brain, both are instrumental in constructing biologically plausible artificial intelligence systems, we introduce the shared dual-memory Transformers (SDMTR)—a model that builds upon Transformers. The proposed approach includes the shared long-term memory and workspace with finite capacity in which different specialized modules compete to write information. Later, crucial information from shared workspace is inscribed into long-term memory through outer product attention mechanism to reduce information conflict and build a knowledge reservoir, thereby facilitating subsequent inference, learning and problem-solving. We apply SDMTR to multi-modality question-answering and reasoning challenges, including text-based bAbI-20k, visual Sort-of-CLEVR and triangle relations detection tasks. The results demonstrate that our SDMTR significantly outperforms the vanilla Transformer and its recent improvements. Additionally, visualiza-

tion analyses indicate that the presence of memory positively correlates with model effectiveness on inference tasks. This research provides novel insights and empirical support to advance biologically plausible deep learning frameworks.

1 INTRODUCTION

The conventional deep neural networks confront a fundamental issue in their tendency to employ a monolithic architecture for the consistent processing of each input sample. For instance, Transformers utilize pairwise attention to establish correlations among disparate positions of inputs (Vaswani et al., 2017; Dosovitskiy et al., 2021), which leads to an overemphasis on information processing and a departure from biological plausibility. One potential solution is to transition toward structuring, modular coordination and sparse interactions, which has exhibited advantages in enhancing model performance and learning efficiency (Minsky, 1987; Brooks, 1991; Greff et al., 2020; Goyal et al., 2022).

The Global Workspace Theory (GWT) (Baars, 1993; Dehaene et al., 1998; VanRullen and Kanai, 2021; Juliani et al., 2022a) and long-term memory (LTM) system represent pivotal concepts in cognitive neuroscience, providing forward-looking guidance in the development of artificial intelligence (AI) systems that adhere more closely to biological norms. Modularization and module coordination are congruent with GWT, wherein multiple specialized modules compete

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

¹Corresponding author: linjie@uestc.edu.cn

to write information into a shared workspace (SW) with constrained capacity that allows cognitive information to be selected, maintained and shared to support advanced cognitive functions such as reasoning and planning. The shared workspace can be likened to a working memory (Baddeley and Hitch, 1974), as it resides within the prefrontal cortex, a region responsible for tasks like abstract rule learning and planning execution (Duncan et al., 1996; O’Reilly and Frank, 2006). While long-term memory is accountable for the persistent retention and retrieval of data, comprising numerous cerebral regions, particularly the hippocampus, which holds paramount role in the formation of new memories (Squire, 1992). These two types of memory play distinct yet pivotal roles in cognition processes, interacting with each other to support holistic cognitive abilities. Although efforts such as Jaegle et al. (2021a,b); Goyal et al. (2022) are valuable initial attempts of a workspace-like method to improve functionality in AI, there are still some way from building biologically plausible AI systems since they all disregard the long-term memory.

Taking inspiration from the GWT and the LTM system, we propose SDMTR— a variant of Transformers, characterized by a shared workspace and a LTM component. Within SDMTR, there are two crucial aspects: the renewal of the two components, along with the broadcast and retrieval of information. Concretely, different specialized modules compete for write access to the shared workspace, where essential information undergo updates in LTM via outer product attention to minimize data conflicts and form knowledge deposits that provide priors for diverse inference tasks. Retrieved memories from LTM can adjust and supplement specialized modules that receive modifications via the broadcast of workspace content, leading to an improvement in performance and generalization.

We have evaluated our SDMTR on a wide range of multimodal question-answering and reasoning tasks, including text-based bAbI-20k, visual Sort-of-CLEVR and Triangle datasets. The results indicate that our SDMTR outperforms vanilla transformers and their recent advancements on accuracy and convergence speed.

2 RELATED WORK

2.1 Recurrent and Memory

Owing to its feedforward nature, Transformers (Vaswani et al., 2017) is unsuitable for state-tracking (Fan et al., 2020), making it less proficient in intricate relational reasoning tasks that demand extensive context and long dependencies. To over-

come this constraint, some researchers have sought to introduce the concepts of recurrent and memory into Transformers to enhance its inductive bias. For example, the Universal Transformer (Dehghani et al., 2018) combines the parallelizability and global receptive field of feed-forward sequence models like Transformers with the recurrent inductive bias of RNNs. In Transformer-XL (Dai et al., 2019) and its successors, including Compressive Transformer (Rae et al., 2019), RMT (Bulatov et al., 2022) and Scaling Transformer (Bulatov et al., 2023), the hidden states from the prior segment are preserved as a memory that is used to augment the current segment, creating a recurrent connection across segments. Unlike Transformer-XL, which abandons past activations as it moves across segments, the Compressive Transformer retain a fine-grained memory of past segment activations. Moreover, Block-Recurrent Transformers (Hutchins et al., 2022) makes use of self-attention and cross-attention to perform a recurrent operation over a large set of state vectors and tokens.

2.2 Sparse and Global Representation

Sparse attention and global representation are two key avenues for improvements of Transformers, which not only reduce computational complexity but also introduce priors from inputs for ameliorated generalization. Here, global representations act as a form of model memory that learns to gather information from input sequence tokens. For instance, Set Transformers (Lee et al., 2019) and Luna (Ma et al., 2021) utilize multiple trainable global nodes to condense input information into a compressed memory that the inputs attend to. Sparse Transformer (Child et al., 2019) combines factorized attention mechanisms with block local attention as well as global attention, where global nodes come from fixed positions in the input sequence. Furthermore, ETC (Ainslie et al., 2020) and Longformer (Beltagy et al., 2020) can be considered as extensions of the Sparse Transformer, both of which introduce global node attention. Star-Transformer (Guo et al., 2019) proposes a sparse pattern that forms a star-shaped graph among nodes with a central global node.

2.3 Transformers with GWT

In addition, there is a growing convergence between recent Transformer advancements and cognitive science theories. The Global Workspace Theory (GWT) on consciousness stands as a widely referenced theory that has enjoyed some of the most lasting influence (Baars, 1993; Mashour et al., 2020), due thanks both to its elaborations over time (Dehaene et al., 1998; Dehaene and Changeux, 2005), as well as the

empirical evidence collected which supports the theory (Dehaene and Changeux, 2011; Van Vugt et al., 2018). For instance, inspired by GWT, Goyal et al. (2022) replace pairwise self-attention in Transformers with a shared workspace through which communication among different input tokens takes place but due to limits on the communication bandwidth, input tokens must compete for access. Further, Sun et al. (2023) combine GWT with associative memory to propose an AiT model—a variant of Transformers with a global workspace layer added behind the feed-forward layer. Besides manual design to reflect the expected attributes of GWT in the model, Juliani et al. (2022b) explored how already widely used machine learning architectures might be compatible with the theoretical requirements of GWT. They argue that Perceiver (Jaegle et al., 2021b) and its variant, Perceiver IO (Jaegle et al., 2021a), also satisfy the criteria of GWT, despite being developed in a separate context with unrelated goals in mind. Our work builds upon Goyal et al. (2022), but distinguishes itself by introducing long-term memory to provide informative priors for updates to the current hidden states. This proposed method can be described as a combination of sparsity, global representation and explicit memory.

3 METHODOLOGY

3.1 Overview

This section provides an overview of the core blocks of Shared Dual Memory Transformers (SDMTR), which consist of four essential phases: the restricted write access and information broadcast in the workspace, as well as the updates and retrieval in the long-term memory, as illustrated in Figure 1. The SDMTR model, in its entirety, replaces the pairwise self-attention of vanilla Transformers with these four steps, where both the workspace and the LTM components are globally shared across all layers. Through end-to-end training, the shared workspace stores a set of priors that are gradually learned from inputs in LTM via outer product attention, which contributes to effective knowledge correlation and precipitation, providing corrections and supplements for subsequent patches update. A more elaborate description of this method is presented as follows.

3.2 Memory Writing

3.2.1 Constrained Writing to the Workspace

We opt for matrix $SW \in \mathbb{R}^{N \times D_m}$ as a form of shared workspace that is initialized randomly. Before the initial writing into the shared workspace, the input undergoes embedding and positional encoding to ob-

tain distinct specialized modules (entity representations for each position), denoted as $\mathcal{E}_0 = [x_{class}; x_p^1 \mathcal{E}; x_p^2 \mathcal{E}; \dots; x_p^T \mathcal{E}] + \mathcal{E}_{pos}$, where $\mathcal{E}_0 \in \mathbb{R}^{T \times D}$. x_{class} is a learnable embedding to the sequence of embedded tokens (similar to BERT’s [class] token). x_p^i represents the i -th patch or token (T patches or tokens in total). \mathcal{E} is a learnable embedding matrix and $\mathcal{E}_{pos} \in \mathbb{R}^{(T+1) \times D}$ is positional embedding. Subsequently, specialized modules compete to write into the shared workspace, whose contents are updated in the context of new information.

To implement the competition, we utilize a multi-headed sparse cross-attention mechanism (MSCA), which is analogous to the multi-headed attention used in Transformers, but with two distinctive features: (i) it requires separate sources for query (Q) and key (K) and (ii) it introduces a sparsity-inducing operation on the attention weight matrix. Concretely, SW^{l-1} , the content of the shared workspace at layer $l-1$, acts as the query, which is matched with the key \mathcal{E}_{t-1}^l , forming attention weights $\mathcal{A}_{sw,e} \leftarrow \text{softmax} \left(\frac{SW^{l-1} \mathcal{W}^Q (\mathcal{E}_{t-1}^l \mathcal{W}^K)^T}{\sqrt{d^k}} \right)$. Then we apply a top- k softmax (Ke et al., 2018) to choose a fixed number of patches authorized to write in the shared workspace, yielding the updated shared workspace $\widetilde{SW}^l \leftarrow \mathcal{A}_{sw,e}^* (\mathcal{E}_{t-1}^l \mathcal{W}^V)$. Here, $\mathcal{A}_{sw,e}^*$ denotes the post-softmax score matrix, obtained by constructing a set $\mathcal{H}t$ containing the indices of the k selected patches with the top- k largest values of $\mathcal{A}_{sw,e}$, as shown in Eq. 1, for all $n \in 1, \dots, N, t \in 1, \dots, T$. This entire selection process can be summarized in Eq. 2.

$$\mathcal{A}_{sw,e}^* = \begin{cases} \mathcal{A}_{[n,t]}, & t \in \mathcal{H}t, \\ 0, & t \notin \mathcal{H}t \end{cases} \quad (1)$$

$$\widetilde{SW}^l = \text{MSCA} \left(SW^{l-1} \mathcal{W}^Q, \mathcal{E}_{t-1}^l \mathcal{W}^K, \mathcal{E}_{t-1}^l \mathcal{W}^V \right) \quad (2)$$

Ultimately, SW^l is generated through residual connections, normalization and feedforward layers, which then goes through the gating mechanism to optimize the contents of the workspace, as depicted in Eq. 3-6, considering it is necessary to erase or reduce the data stored previously, making way for new information.

$$\widehat{SW}^l = \text{LN}(\widetilde{SW}^l + \widetilde{SW}^{l-1}) \quad (3)$$

$$\widehat{SW}_i^l = \text{ReLU}(\text{MLP}_i(\widehat{SW}_{i-1}^l)), i \in \{1, \dots, k\} \quad (4)$$

$$\widehat{SW}_{new}^l = \text{LN}(\widehat{SW}^{l-1} + \widehat{SW}_k^l) \quad (5)$$

$$SW^l = \mathcal{F}_t(SW^{l-1}, \mathcal{E}_{t-1}^l) \odot \widehat{SW}^{l-1} + \mathcal{I}_t(SW^{l-1}, \mathcal{E}_{t-1}^l) \odot \widehat{SW}_{new}^l \quad (6)$$

Here, \mathcal{I}_t and \mathcal{F}_t indicate the input and forget gates respectively, as proposed in RMC (Santoro et al., 2018). Further details can be found in Appendix B.1.

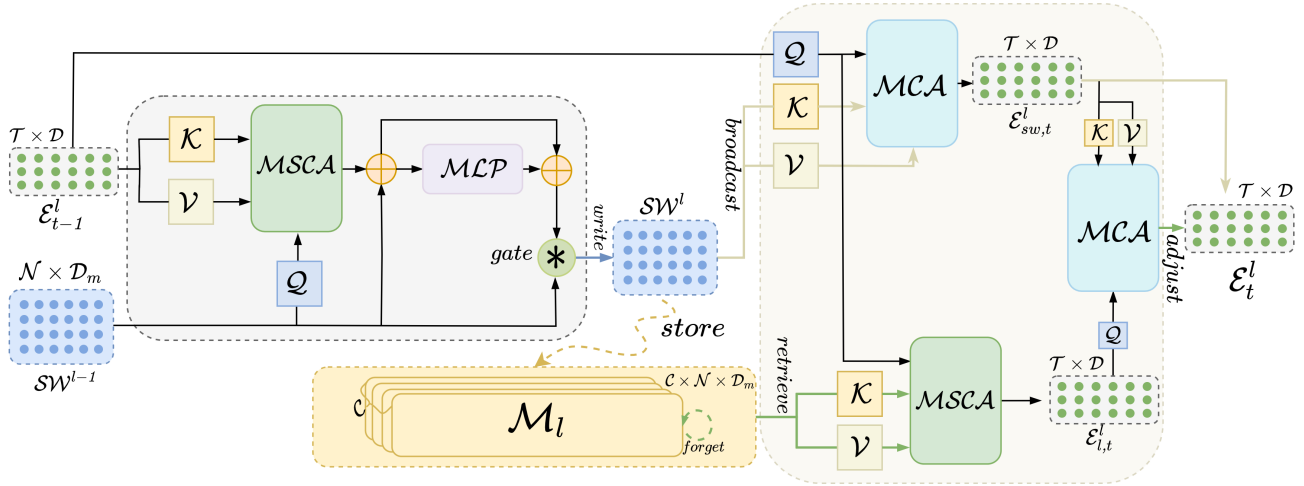


Figure 1: An Overview of the SDMTR core layer. The update module facilitates a competition for write access to shared workspace SW^{l-1} by the input \mathcal{E}_{t-1}^l from the previous computation step $t-1$ in current layer l , yielding the updated SW^l , whose vital information are stored into shared \mathcal{M}_l via outer product attention. The contents of SW^l are then broadcast to specialized locations, fine-tuned by retrieval memories from \mathcal{M}_l to obtain \mathcal{E}_t^l for the next stage t in the current layer l .

3.2.2 Update for Long-term Memory

Our representation of long-term memory \mathcal{M}_l takes the form of a 3D structure in $\mathbb{R}^{C \times N \times D_m}$, with C signifying memory fragments, which is biologically plausible (Marr and Thach, 1991). Utilizing outer product attention (OPA), we transfer the contents of the workspace into long-term memory, which boasts a larger capacity for extensive information that can provide a more substantial set of priors for inference. OPA—a natural extension of the query-key-value dot product attention (DPA) (Vaswani et al., 2017), offers two main benefits: (i) a higher-order representational capacity, more powerful than that of DPA and (ii) more complex interactions between different lines of data sources, which contribute to information precipitation and memory reinforcement.

As illustrated in Eq. 7, we perform the OPA between the updated SW^l and the previous layer LTM \mathcal{M}_l^{l-1} to merge novel vital information into the current LTM, where \odot is element-wise multiplication, \otimes is outer product and \mathcal{G} is chosen as element-wise tanh function. Finally, the updated \mathcal{M}_l^l is obtained by residual connection and normalization in Eq. 8.

$$\begin{aligned} \widetilde{\mathcal{M}}_l^l &= \mathcal{A}^{\otimes}(\mathcal{M}_l^{l-1}, SW^l, SW^l) = \\ &= \sum_{i=1}^{n_{kv}} \mathcal{G} \left(\frac{\mathcal{M}_{l,i}^{l-1} \odot SW_i^l}{\sqrt{d_k}} \right) \otimes SW_i^l \end{aligned} \quad (7)$$

$$\mathcal{M}_l^l = LN \left(\widetilde{\mathcal{M}}_l^l + \mathcal{M}_l^{l-1} \right) \quad (8)$$

3.3 Memory Reading

3.3.1 Information Broadcast from Workspace

Firstly, each specialized module refreshes its state using the information broadcast from the shared workspace. This consolidation is conducted through the multi-head cross-attention mechanism (MCA), which is equivalent to the MSCA but without top-k softmax. In this context, all the patches create queries, while the updated SW^l functions as both keys and values, resulting in the creation of new specialized representations $\mathcal{E}_{sw,t}^l$, as shown in Eq. 9, where \widetilde{W}^Q , \widetilde{W}^K and \widetilde{W}^V stand for weight matrices.

$$\mathcal{E}_{sw,t}^l = MCA \left(\mathcal{E}_{t-1}^l \widetilde{W}^Q, SW^l \widetilde{W}^K, SW^l \widetilde{W}^V \right) \quad (9)$$

3.3.2 Retrieval from Long-term Memory

We contend that it is inadequate for complex inference tasks relying solely on information from the shared workspace broadcast, since it resembles a working memory with limited capacity, capable of temporarily storing and processing minimal information. Considering the practical abilities of humans to focus on critical details of the ongoing task and leverage their extensive knowledge and experience to navigate complex situations, it intuitively implies that retrieving relevant information from long-term memory is valuable for intricate reasoning tasks. Therefore, we again utilize content-based addressing MSCA to retrieve pertinent priors from long-term memory for refinement and supplementation of specialized modules. The current

input serves as Q and the updated long-term memory as K and V, resulting in an instructive representation $\mathcal{E}_{l,t}^l$, as denoted in Eq. 10

$$\mathcal{E}_{l,t}^l = \mathcal{MSCA} \left(\mathcal{E}_{t-1}^l \widehat{\mathcal{W}}^Q, \mathcal{M}_l^l \widehat{\mathcal{W}}^K, \mathcal{M}_l^l \widehat{\mathcal{W}}^V \right) \quad (10)$$

Subsequently, the priors $\mathcal{E}_{l,t}^l$ come into play to refine and enhance the understanding $\mathcal{E}_{sw,t}^l$ via the MCA mechanism, resulting in $\mathcal{E}_{swl,t}^l$. Afterward, we perform a weighted combination of $\mathcal{E}_{sw,t}^l$ and $\mathcal{E}_{swl,t}^l$ using the hyper-parameter β to obtain the ultimate states \mathcal{E}_t^l , as outlined below. This process aims to extract more profound and valuable insights, enabling advanced decision-making and reasoning.

$$\mathcal{E}_{swl,t}^l = \mathcal{MCA} \left(\mathcal{E}_{l,t}^l \bar{W}^Q, \mathcal{E}_{sw,t}^l \bar{W}^K, \mathcal{E}_{sw,t}^l \bar{W}^V \right) \quad (11)$$

$$\mathcal{E}_t^l = \beta \mathcal{E}_{sw,t}^l + (1 - \beta) \mathcal{E}_{swl,t}^l \quad (12)$$

4 EXPERIMENTS

In this section, we evaluate our SDMTR on various relation reasoning tasks, including visual and text-based question-answering, as well as detecting equilateral triangles. The experimental datasets encompass bAbI, Sort-of-CLEVR and Triangle datasets. Furthermore, we conduct ablation experiments on three datasets in Section 4.4, and Section 4.5 provides some insights into interpretability. Detailed parameter settings for each experiment are shown in Appendix B.2.

4.1 Relational Reasoning : Sort-of-CLEVR

Sort-of-CLEVR (Santoro et al., 2017) is a dataset similar to CLEVR, designed specifically for visual relational reasoning tasks, which includes answering relational and non-relational problems. Considering the restricted answer choices, this task is classified as a classification task. Each 2D image in this dataset measures 75×75 pixels and is accompanied by six geometric shapes, which are randomly placed and can be colored in one of six available colors, with a choice of two shapes. Per image is associated with 10 unary questions, i.e. non-relational questions, as well as 10 binary and 10 ternary questions, both of which are relational problems (details in Appendix E.2). Following ViT (Dosovitskiy et al., 2020), a sequence of fixed-size patches for each image are generated, which are then concatenated with the corresponding question embedding as inputs into our SDMTR, in line with Goyal et al. (2022).

Baselines For this task we evaluated our SDMTR with the following five baselines: Transformers [TR] (Vaswani et al., 2017), Set transformer [ISAB]: Transformers where self-attention is replaced by ISAB

module (Lee et al., 2019), Transformers with Shared Workspace with top-k competition [TR+HSW] (Goyal et al., 2022), High Capacity Transformers [TR+HC]: Same as TR but with different parameters across layers and Associative Transformer [AiT] (Sun et al., 2023): Transformers with the introduction of a global workspace layer following the original feedforward layer layer.

The test accuracy curves for 200 epochs of all models are illustrated in Fig. 2. It’s obvious that our SDMTR stands out compared to all other baselines, exhibiting superior performance in both relational and non-relational tasks, with faster convergence and higher accuracy. Conversely, the TR+HSW with only the workspace and the AiT with the added global workspace layer excel in non-relational problems but fall short in dealing with relational problems. We suspect this is because unary problems involve handling rare information related to individual objects, and the limited memory slots in the global workspace, designed for storing and processing current data, can easily manage these tasks. Nevertheless, for relational problems, solutions often entail multi-step reasoning, like object attribute extraction followed by relational analysis. The high-level information and complex relations stored in LTM can be accessed as priors, going beyond reliance on current inputs, thus facilitating a more thorough comprehension and resolution of relational issues.

4.2 Text-based QA : bAbI

The BAbI dataset is a textual question answering benchmark (Weston et al., 2015) that is widely utilized to assess the memory and reasoning capabilities of models, including Memory-augmented Neural Networks (MANNs), Recurrent Neural Networks (RNNs) and Networks based on attention mechanisms. This dataset consists of 20 tasks, each of which is subdivided into training, validation and test datasets, with 9k, 1k and 1k questions respectively, related to logical deduction, counting, pathfinding and induction. Each task is presented in the form of short stories or text passages, comprising narratives, questions, answers and supporting facts. For example, the facts “Brian is a lion”, “Bernhard is white” and “Brian is a lion” support the question “What color is Brian?” (answer: ‘white’).

Following Le et al. (2020b), each story is transformed into a sentence-level sequence, which serves as the input for our SDMTR model (details in Appendix E.1). We use normal supervised training to jointly train the SDMTR for all tasks and report the results in Table 1. A model can be deemed successful in task execution if its performance surpasses 95%, and in this

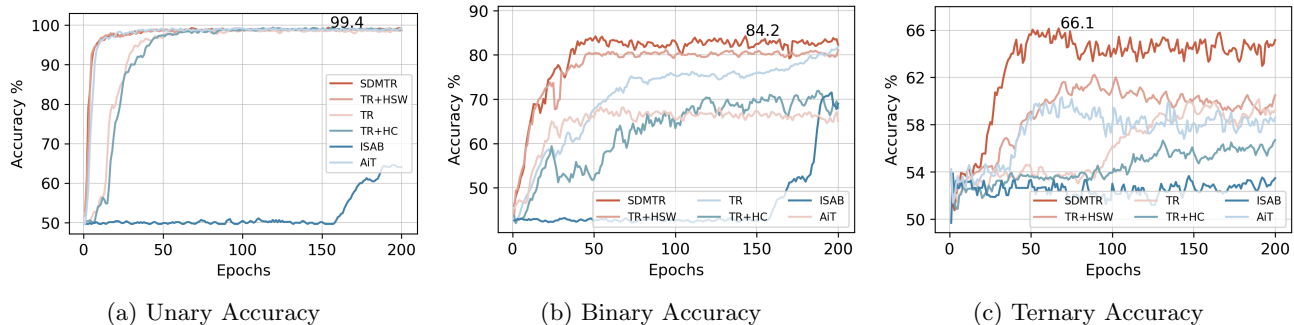


Figure 2: Test accuracy vs epochs for the Sort-of-CLEVR task.

Table 1: Test error rates: mean \pm std. (in %) on the 20 bAbI tasks for models jointly trained with 10k examples and best error over 10 runs. \dagger is reported from Dehghani et al. (2018)

Model	Error	
	Mean	Best
LSTM (Hochreiter and Schmidhuber, 1997)	27.3 \pm 0.8	25.2
TR \dagger (Vaswani et al., 2017)	22.1	N/A
DNC (Graves et al., 2016)	12.8 \pm 4.7	3.8
H-Mem (Limbacher and Legenstein, 2020)	10.8	N/A
NUTM (Le et al., 2020a)	5.6 \pm 1.9	3.3
MemNet (Dou and Principe, 2023)	5.6	N/A
TR+HSW (Goyal et al., 2022)	3.6 \pm 0.46	3.25
SDMTR (ours)	3.15\pm0.24	2.95

case, SDMTR succeeded in all 19 tasks (more in Appendix D). The significant decline in error rates for TR, TR+HSW and our SDMTR provides evidence of the effectiveness of global workspace and long-term memory in inference tasks. In addition, MANNs like DNC and NUTM exhibit relatively higher error rates, which we postulate may result from their lack of explicit support for relations learning.

4.3 Detecting Equilateral Triangles

Our objective in this task is to determine the presence of an equilateral triangle, which consists of three point clusters randomly positioned within a 64×64 sized image (Ahmad and Omohundro, 2009). The condition for forming an equilateral triangle is that the mid-points within these clusters maintain equal distances from each other. Given that the answer is yes or no, this is a binary classification problem. In order to feed images into our SDMTR model, we split each image into patches of size 16×16 , which are then employed as different input positions for SDMTR, akin to the approach proposed in ViT (Dosovitskiy et al., 2020).

The results are reported in Table 2, where our SDMTR (2 layers and 4 heads) achieves an impressive accu-

racy of 98.1%, which not only outperforms the vanilla Transformers (TR) by 8.3% but also surpasses the TR+HSW and Perceiver based on the GWT theory by 1.43% and 1.36%, respectively. The improvement in TR+HSW and Perceiver compared to TR highlights the successful application of GWT. Furthermore, the further progress observed in our SDMTR demonstrates that the priors retrieved from long-term memory do have a positive effect on the specialized modules, which are updated by the information broadcast from shared workspace. Our hypothesis regarding SDMTR’s effectiveness in grasping and memorizing spatial relations among different clusters, including their relative positions and distances, is attributed to its utilization of a global shared workspace and long-term memory, which allows it to retain essential data about each cluster point. In addition, limited capacity of workspace compels the model to selectively record critical information into memories, which is consistent with the inherent sparsity of the task. Here, STR denotes Transformers with sparse factorizations of the attention matrix (Child et al., 2019), [SDMTR+S] is a variant of the SDMTR without top-k sparsity, \dagger is reported from Sun et al. (2023) and other baselines with the same configuration as SDMTR are detailed in experiment 4.1.

4.4 Ablation Studies

In this section, we carry out ablation studies to explore how the model’s performance is impacted by factors, including model size, memory capacity, memory persistence (global sharing), constrained access during the writing process, as well as the adjustment of priors derived from LTM. To tackle these questions, we run our $SDMTR_s$ and $SDMTR_m$ on various combinations of N , M and k , where N and M control the capacity of SW and LTM, respectively, k determines the size of top-k. For equilateral triangle detection, $SDMTR_s$ has $l = 2$, $h = 4$, and $SDMTR_m$ has $l = 4$, $h = 4$, while for the other two tasks, $SDMTR_s$ has $l = 4$, $h = 4$, and $SDMTR_m$ has $l = 8$, $h = 8$. Here, l

Table 2: Performance comparisons for equilateral triangle detection between our SDMTR and other Transformer baseline models.

Models	STR	ISAB	TR	TR+HSW	AiT	Perceiver [†]	SDMTR	SDMTR+S
Acc (%)	61.12	60.93	89.84	96.71	98.05	96.78	98.14	97.67

Table 3: Results of ablation studies on memory properties across three datasets.

Model	N	C	Top- k	Sort-of-CLEVR				bAbI		Triangle	
				Params	Unary%	Binary%	Ternary%	Params	Err%	Params	Acc%
$SDMTR_s$	6	3	5	2.44M	99.12	79.62	61.85	2.43M	3.14	2.42M	97.94
	8	5	5	2.54M	99.40	84.17	63.49	2.53M	2.95	2.51M	98.14
	8	5	7	2.54M	99.35	82.07	66.13	2.53M	2.97	2.51M	98.09
	10	7	9	2.63M	99.17	80.08	60.07	2.62M	3.06	2.61M	97.90
$SDMTR_m$	6	3	5	2.63M	99.38	80.12	62.44	2.62M	2.99	2.61M	98.16
	8	5	5	2.65M	99.41	85.06	64.18	2.64M	2.82	2.62M	98.37
	8	5	7	2.65M	99.36	83.01	66.82	2.64M	2.84	2.62M	98.30
	10	7	9	2.76M	99.25	80.99	60.89	2.71M	2.91	2.70M	98.13
$SDMTR_{s w/o_1}$	8	5	5	10.16M	99.11	78.19	59.78	9.64M	3.18	4.82M	95.27
$SDMTR_{s w/o_2}$	8	5	5	2.48M	99.13	79.06	61.35	2.47M	3.24	2.46M	96.16
$SDMTR_s$	8	5	<i>soft</i>	2.54M	99.15	78.34	61.16	2.52M	3.05	2.51M	97.25

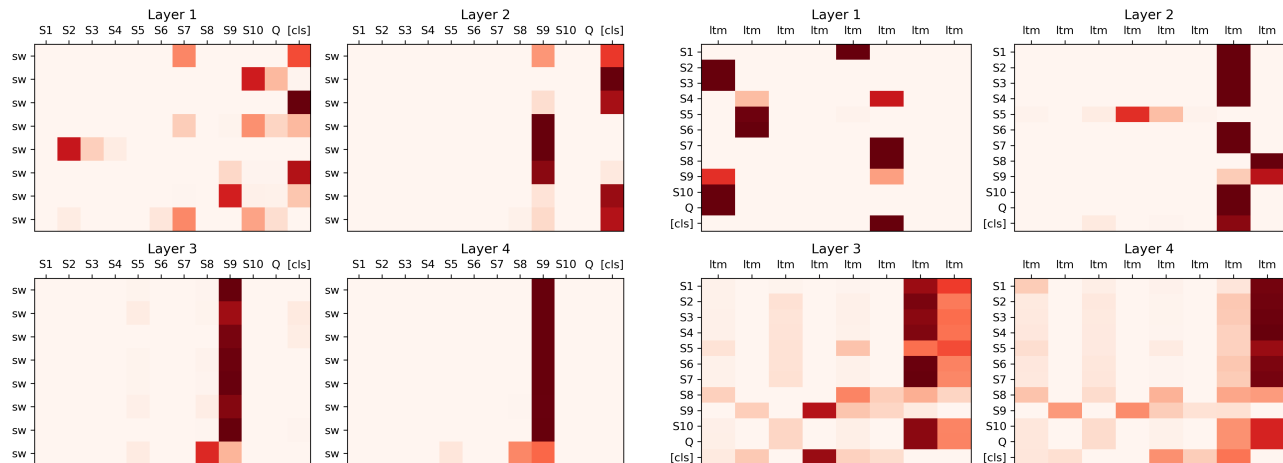
and h denote the number of layers and heads in the model, respectively (more in Appendix B.2). The results are reported in Table 3, where $SDMTR_{s w/o_1}$ denotes $SDMTR_s$ without memory sharing among all layers, while $SDMTR_{s w/o_2}$ indicates that priors retrieved from LTM are directly aggregated with information from WM via β without correction step. *soft* is a standard soft competition mode, not a top- k strategy.

The following four points encapsulate our key discoveries. Firstly, a larger memory capacity doesn’t guarantee better performance, since the larger the capacity, the harder the matrix is to stabilize and converge. The most favorable outcomes are commonly observed at $N = 8$ and $M = 5$, which is consistent with findings in cognitive neuroscience, where Miller’s law states that the number of information units that humans can handle simultaneously is limited to around 7 ± 2 . Notably, holding separate memory for each layer brings another considerable drawback: a linear growth in trainable parameters as the number of layer increases, e.g. a fourfold increase in a four layers model. Secondly, the absence of memory persistence leads to a substantial drop in accuracy for the binary and ternary relational inference tasks, with a respective decline of 5.98% and 6.35%, compared to the best case with global sharing (For accuracy curves, refer to Appendix A). Thirdly, the model’s performance exhibits considerable sensi-

tivity to different values of k , with $k = \{5, 7, \text{soft}\}$ resulting in two distinct phenomena across the three tasks. More precisely, except for the ternary problems in the Sort-of-CLEVR dataset, which achieve their highest accuracy at $k = 7$, the other tasks perform optimally at $k = 5$. We conjecture that this disparity could be ascribed to the increased demand for data storage in ternary problems, hence necessitating a slightly higher value for k . The fourth point concerns the influence of priors stored in LTM. Under the same conditions, the lack of LTM guidance leads to a decrease in accuracy for the three types of Sort-of-CLEVR tasks, with declines of 0.27% (unary problems), 5.11% (binary problems) and 4.78% (ternary problems). This unfavorable trend extends to two other tasks as well, with a 0.29% rise in the error rate of the bAbI dataset and a 1.98% drop in accuracy for the triangle dataset. Taken together, these results demonstrate the instructive support provided by prior knowledge accumulated in LTM for relational inference.

4.5 Visualization

In this section, we provide some interpretability about this work. Using the bAbI dataset as a case study, we visualized two specific attention patterns related to the competitive writing in the shared workspace and



(a) Attention pattern on writing with restricted access to the shared workspace.

(b) Attention patterns for retrieving information from long-term memory.

Figure 3: Visualization of two distinct attention patterns across different layers in the SDMTR model.

the retrieval process from long-term memory, as depicted in Fig. 3. Here, S1 to S10 represent ten distinct sentence embeddings that collectively provide the contextual support for answering the question. Q signifies the question embedding, and CLS is the classification head. The saturation of color conveys the extent of correlation between the current input and what is stored in a shared workspace or long-term memory, where the deeper the color, the stronger the relevance.

The heatmaps for both cases demonstrate opposite trends as the layer count rises (see Appendix C for details). Specifically, in Fig. 3a, as model layers increase, the heatmap shows a reduction in the number of colored blocks and an intensification of color within the remaining blocks during the constrained writing process. This signifies a progressively clearer and more distinct focus area for the model, where the shared workspace that acts as a latent representation can proficiently extract the most crucial information from the input sequence. In Fig. 3b, the colored areas steadily expand as layers increase, and the heat map tends to be stable and convergent, which indicates that the correlation between the current input and the content in long-term memory gradually becomes higher, that is, more information about the input sequence can be retrieved from long-term memory. This provides favorable evidence supporting the potential for extensive knowledge consolidation in long-term memory, allowing it to offer informative priors for the update of current hidden state.

5 CONCLUSION

We have proposed a SDMTR model that builds upon vanilla Transformers, drawing inspiration from global workspace theory and long-term memory system in cognitive neuroscience. In SDMTR, the pairwise self-attention layer found in vanilla Transformers is replaced with a layer that comprises a workspace and a long-term memory, both shared globally across all layers. Communication bottleneck and outer-product attention are utilized for distinct updates to the workspace and long-term memory, while content-based addressing serves as the shared technique for the broadcasting within the workspace and retrieval from long-term memory. The ability of SDMTR to reason in complex scenarios is validated through a suite of diverse tasks including the bAbI and Sort-of-CLEVR question-answering, as well as triangle relations detection. In all cases, our model demonstrates strong performance, confirming its effectiveness of possessing both workspace and long-term memory in one model. In future work, we look forward to exploring how to better integrate GWT theory, long-term memory and other cognitive components with modern deep neural networks and in addition, we plan to broaden the scope of our research to a wider array of multimodal reasoning tasks, both of which we consider to be crucial leaps towards building a strong AI.

Acknowledgements

This work was supported by Sichuan Province Science and Technology Support Program, No.:2021YFN0117.

References

- Ahmad, S. and Omohundro, S. M. (2009). Equilateral triangles: A challenge for connectionist vision.
- Ainslie, J., Ontanon, S., Alberti, C., Cvicek, V., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q., and Yang, L. (2020). Etc: Encoding long and structured inputs in transformers. *arXiv preprint arXiv:2004.08483*.
- Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.
- Baddeley, A. D. and Hitch, G. (1974). Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier.
- Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.
- Bulatov, A., Kuratov, Y., and Burtsev, M. (2022). Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091.
- Bulatov, A., Kuratov, Y., and Burtsev, M. S. (2023). Scaling transformer to 1m tokens and beyond with rmt. *arXiv preprint arXiv:2304.11062*.
- Child, R., Gray, S., Radford, A., and Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- Dehaene, S. and Changeux, J.-P. (2005). Ongoing spontaneous activity controls access to consciousness: a neuronal model for inattentive blindness. *PLoS biology*, 3(5):e141.
- Dehaene, S. and Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227.
- Dehaene, S., Kerszberg, M., and Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the national Academy of Sciences*, 95(24):14529–14534.
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Lukasz Kaiser (2018). Universal transformers.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Dou, R. and Principe, J. (2023). Universal recurrent event memories for streaming data. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.
- Duncan, J., Emslie, H., Williams, P., Johnson, R., and Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive psychology*, 30(3):257–303.
- Fan, A., Lavril, T., Grave, E., Joulin, A., and Sukhbaatar, S. (2020). Addressing some limitations of transformers with feedback memory. *arXiv preprint arXiv:2002.09402*.
- Goyal, A., Didolkar, A. R., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M. C., and Bengio, Y. (2022). Coordination among neural modules through a shared global workspace. In *International Conference on Learning Representations*.
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476.
- Greff, K., Van Steenkiste, S., and Schmidhuber, J. (2020). On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.
- Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., and Zhang, Z. (2019). Star-transformer. *arXiv preprint arXiv:1902.09113*.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Hutchins, D., Schlag, I., Wu, Y., Dyer, E., and Neyshabur, B. (2022). Block-recurrent transformers. *Advances in Neural Information Processing Systems*, 35:33248–33261.
- Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., et al. (2021a). Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*.
- Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021b). Perceiver:

- General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.
- Juliani, A., Arulkumaran, K., Sasai, S., and Kanai, R. (2022a). On the link between conscious function and general intelligence in humans and machines. *arXiv preprint arXiv:2204.05133*.
- Juliani, A., Kanai, R., and Sasai, S. S. (2022b). The perceiver architecture is a functional global workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.
- Ke, N. R., ALIAS PARTH GOYAL, A. G., Bilaniuk, O., Binas, J., Mozer, M. C., Pal, C., and Bengio, Y. (2018). Sparse attentive backtracking: Temporal credit assignment through reminding. *Advances in neural information processing systems*, 31.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Le, H., Tran, T., and Venkatesh, S. (2020a). Neural stored-program memory. In *International Conference on Learning Representations*.
- Le, H., Tran, T., and Venkatesh, S. (2020b). Self-attentive associative memory. In *International Conference on Machine Learning*, pages 5682–5691. PMLR.
- Lee, J., Lee, Y., Kim, J., Kosiosek, A., Choi, S., and Teh, Y. W. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR.
- Limbacher, T. and Legenstein, R. (2020). H-mem: Harnessing synaptic plasticity with hebbian memory networks. *Advances in Neural Information Processing Systems*, 33:21627–21637.
- Ma, X., Kong, X., Wang, S., Zhou, C., May, J., Ma, H., and Zettlemoyer, L. (2021). Luna: Linear unified nested attention. *Advances in Neural Information Processing Systems*, 34:2441–2453.
- Marr, D. and Thach, W. T. (1991). A theory of cerebellar cortex. *From the Retina to the Neocortex: Selected Papers of David Marr*, pages 11–50.
- Mashour, G. A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798.
- Minsky, M. (1987). The society of mind. In *The Personalist Forum*, volume 3, pages 19–32. JSTOR.
- O’Reilly, R. C. and Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2):283–328.
- Rae, J. W., Potapenko, A., Jayakumar, S. M., and Lillicrap, T. P. (2019). Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.
- Santoro, A., Faulkner, R., Raposo, D., Rae, J., Chrzanowski, M., Weber, T., Wierstra, D., Vinyals, O., Pascanu, R., and Lillicrap, T. (2018). Relational recurrent neural networks. *Advances in neural information processing systems*, 31.
- Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., and Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.
- Squire, L. R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, 99(2):195.
- Sun, Y., Ochiai, H., Wu, Z., Lin, S., and Kanai, R. (2023). Associative transformer is a sparse representation learner. *arXiv preprint arXiv:2309.12862*.
- Van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., and Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, 360(6388):537–542.
- VanRullen, R. and Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences*, 44(9):692–704.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Weston, J., Bordes, A., Chopra, S., Rush, A. M., Van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Checklist

- For all models and algorithms presented, check if you include:
 - A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes
 - An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes
 - (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] Yes

2. For any theoretical claim, check if you include:
- (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Not Applicable
 - (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Not Applicable
 - (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes
- (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable
- (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable] Not Applicable
3. For all figures and tables that present empirical results, check if you include:
- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
- (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Yes
 - (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Yes
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Yes
 - (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Not Applicable
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] No
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
- (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Not Applicable
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board

A ADDITIONAL EXPERIMENTAL RESULTS

In this section, we provide a more detailed elaboration of the ablation experiments discussed in Section 4.4 of the main text. Figure 4 and Figure 5 illustrate the variations in error rate and test accuracy on the bAbI dataset and the Sort-of-CLEVR dataset during training iterations, respectively. Here, the SDMTR (4 layers and 4 heads) model represents an improved version of Transformers with a globally shared workspace and long-term memory module, as proposed in this paper. The *SDMTR_NS* ($SDMTR_{w/o1}$) denotes a configuration where the workspace and long-term memory are not globally shared, that is, each layer possesses its independent memory. *SDMTR_NL* ($SDMTR_{w/o2}$) signifies a setup in which content retrieved from long-term memory is directly superimposed on hidden states that have received broadcast information from workspace, without the cross-attention step between the workspace and long-term memory. Soft refers to the use of a soft competition mechanism for workspace writing, as opposed to top-k selection method.

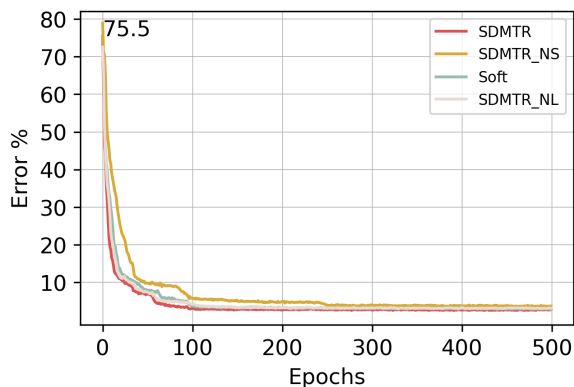


Figure 4: Test error rate vs. training epochs for the bAbI-20k dataset.

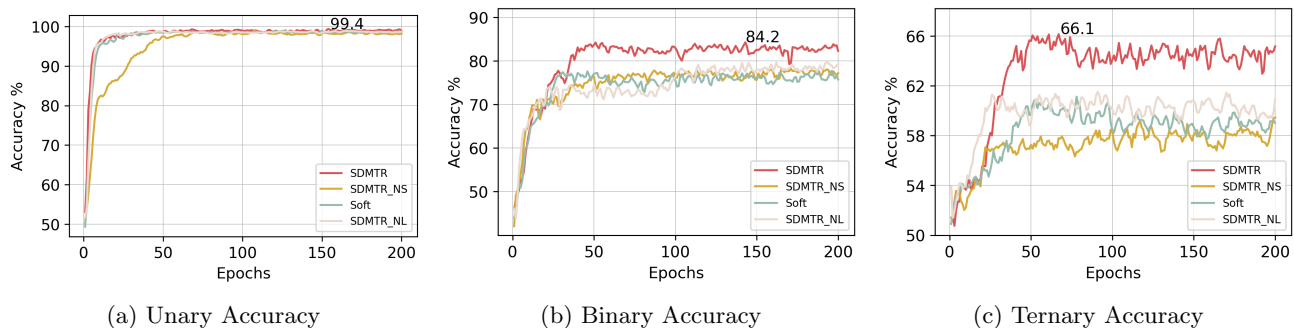


Figure 5: Test accuracy vs. training iterations for the Sort-of-CLEVR task.

We can infer the following discoveries from Figure 4 and Figure 5. Firstly, across diverse tasks in both datasets, the SDMTR consistently attains the highest accuracy, with faster convergence in most cases compared to other models. Secondly, the impact of non-global sharing outweighs the absence of interaction between long-term memory and the workspace, particularly evident in ternary problems and the bAbI task, highlighting the efficacy of the adopted global sharing strategy in complex relation inference. Thirdly, for unary problems in Sort-of-CLEVR dataset, their performances remain relatively on par, which we suspect may be attributed to non-relational tasks typically focus on extracting direct patterns and information from inputs, whereas other ablation models with multiple cross-attention mechanisms can effectively capture these correlations, providing a plausible explanation for their approximate performances.

B EXPERIMENTAL DETAILS

B.1 Gating Mechanism

Drawing inspiration from Santoro et al. (2018), we introduce gating mechanism to update the workspace, with a primary focus on detailed designs for input and forget gates. Let $\mathcal{E}_{t-1}^l = [\mathcal{E}_{t-1,1}^l, \mathcal{E}_{t-1,2}^l, \dots, \mathcal{E}_{t-1,T}^l] \in \mathbb{R}^{T \times D}$ represent the current input sequence, and \mathcal{SW}^{l-1} , $\mathcal{SW}^l \in \mathbb{R}^{N \times D_m}$ denote the prior and updated workspace, respectively. $\widetilde{\mathcal{SW}}_{new}^l$ is an intermediate result as described in Equation 5 in Section 3.2.1 of the main text. The gating mechanism can be formulated as follows:

$$\begin{aligned}\bar{\mathcal{E}} &= \frac{1}{T} \sum_{i=1}^T \text{relu}(\mathcal{E}_i^l \times W^I) \\ K &= \bar{\mathcal{E}} + \tanh(\mathcal{SW}^{l-1}) \times W^F \\ \mathcal{I}_t &= \text{sigmoid}(K + b_i) \\ \mathcal{F}_t &= \text{sigmoid}(K + b_f) \\ \mathcal{SW}^l &= \mathcal{I}_t \times \tanh(\widetilde{\mathcal{SW}}_{new}^l) + \mathcal{F}_t \times \mathcal{SW}^{l-1}\end{aligned}$$

Here, \mathcal{I}_t and \mathcal{F}_t correspond to the input and forget gates of the current calculation step t , with the associated biases b_i and b_f , respectively. W^I and W^F stand for weight matrices. In practice, we configure $b_i = 0$, $b_f = 1$, $D = D_m$.

B.2 Parameter Settings

For visual tasks like Sort-of-CLEVR and Equilateral Triangle, it takes about 8 hours to run 200 epochs on a V100 (24G) GPU. In the case of the bAbI-20k task, it takes about 2 days to jointly train on a V100 (16G) GPU. The detailed hyperparameter settings of our SDMTR on the three datasets are outlined in Table 4, where the Adaptive Moment Estimation (Adam) algorithm proposed by Kingma and Ba (2014) is used to dynamically fine-tune the learning rates during training.

Table 4: The hyperparameter settings of our SDMTR on three datasets

Parameters	Tasks		
	bAbI	Sort-of-CLEVR	Triangle
Top-k	5	5	5
Number of layers	4	4	2
Number of attention heads	4	4	4
Embedding dimensions	128	256	128
Optimizer	Adam	Adam	Adam
Patch size	N/A	15	32
Learning rate	0.0002	0.0001	0.0001
Batch size	64	32	100
Inp Dropout	0.1	0.1	0.1
Seed	1	1	1
Number of memory slots in workspace (N)	8	8	8
Number of long-term memory segments (C)	5	5	5
Size of each working memory slot (D_m)	128	256	128
Number of MLP layers in attention	4	4	5
Memory attention heads	4	4	1
Gate style	'unit'	'unit'	'unit'
Initial β value	0.7	0.75	0.7

C VISUALIZATION ANALYSIS OF ATTENTION PATTERNS

In this part, we present additional examples to elucidate the statements made in Section 4.5 of the main text. Specifically, we illustrate the attention patterns during the competition for workspace update and retrieval from long-term memory for four question-answering instances from the bAbI dataset, as shown in Figure 6 and 7, respectively. In each row, different question-answering examples are represented, where the columns correspond to layers in the SDMTR model. S1- S_{max} represent the embeddings of various background sentences that collectively support the answer to the question. For instance, S1 denotes ‘The pink rectangle is to the left of the triangle’ and S2 denotes ‘The triangle is to the left of the red square’, and so on, up to S_{max} , with the maximum S_{max} varying for each task (in task 1, S_{max} is S10). When there are fewer than S_{max} story sentences, we use placeholder sentences (consisting of all zeros) as filler input. Q represents the question embedding, e.g. ‘Is the pink rectangle to the right of the red square?’, and ‘CLS’ is the classification header.

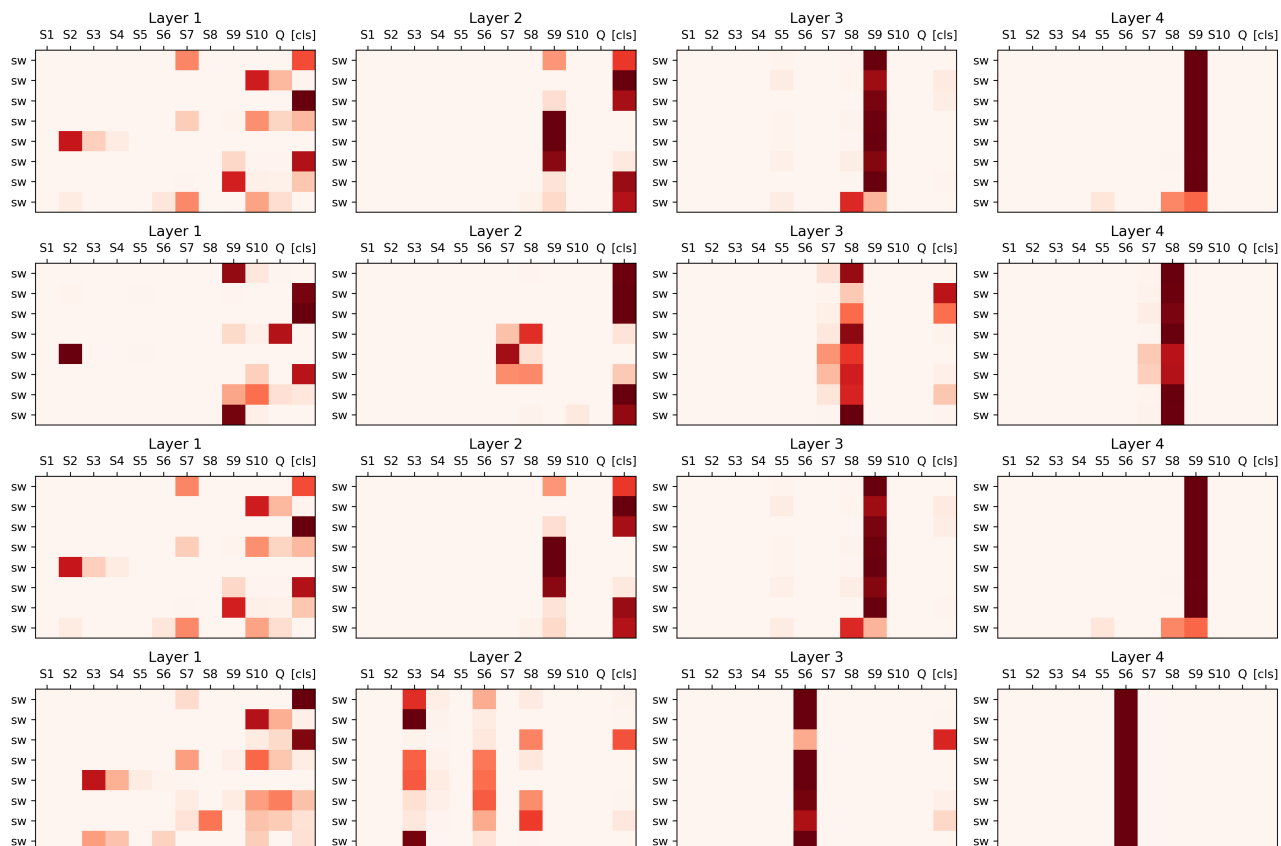


Figure 6: Attention pattern on writing with restricted access to the shared workspace.

D BABI DETAILED RESULTS

The detailed experimental results of the bAbI dataset in Section 4.2 of the main text are shown in Table 5.

E DATASET DESCRIPTION

E.1 BABI

Comprising 20 distinct text-based QA tasks, the bAbI-20k dataset poses various reasoning challenges, ranging from counting to deduction and induction. It is partitioned into training, validation and test datasets, with 9k, 1k and 1k questions, respectively. These tasks are presented in the form of short stories or text passages, including narratives, questions, answers and supporting facts. The narratives introduce entities, actions and

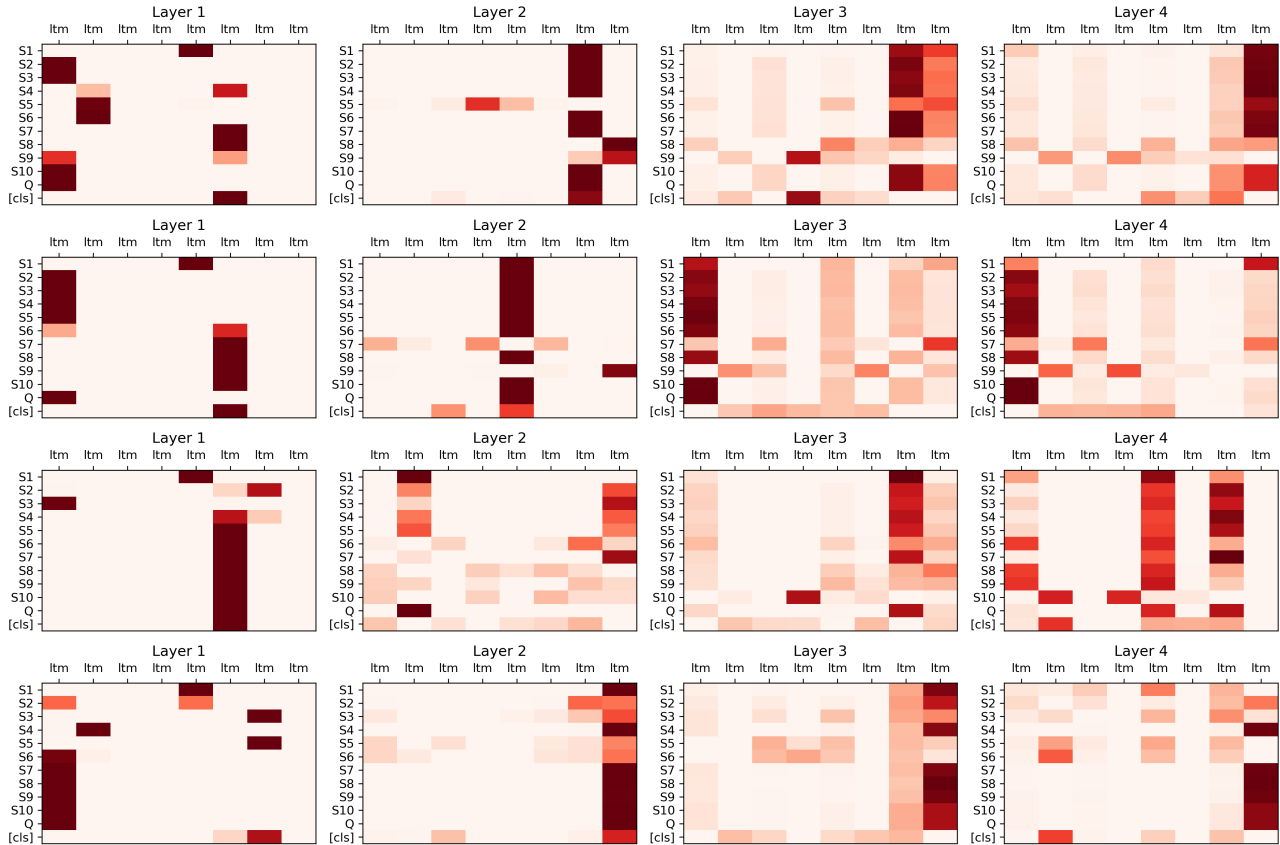


Figure 7: Attention patterns for retrieving information from long-term memory.

contextual information pertinent to the question, while the answer is substantiated by facts from the narratives. Here are four illustrative examples.

For the bAbI task, we employ integer encoding to map all words appearing in the 20 subtasks to a unique integer ID. Each story is represented as a tensor of shape (story_max, sentence_max), where story_max denotes the maximum of sentences involved in any question of a subtask, and sentence_max is the maximum of words across all sentences in that subtask. Any shortfall is padded with zeros. Subsequently, the integer-encoded stories and questions undergo an embedding layer, followed by concatenation, incorporating a cls_token to serve as inputs for our SDMTR.

Task 1: Single Supporting Fact

1. John travelled to the hallway.
 2. Mary journeyed to the bathroom.
 3. Daniel went back to the bathroom.
 4. John moved to the bedroom.
- Q: Where is Mary?
A: bathroom S: 2

Task 7: Counting

1. Sandra went to the bedroom.
 2. Mary went to the office.
 3. Mary took the apple there.
 4. Mary put down the apple.
- Q: How many objects is Mary carrying?
A: none S: 3 4

Task 14: Time Reasoning

1. This morning Mary moved to the kitchen.
 2. This afternoon Mary moved to the cinema.
 3. Yesterday Bill went to the bedroom.
 4. Mary went to the school this evening.
- Q: Where was Mary before the school?
A: cinema S: 2 4

Task 18: Size Reasoning

1. The suitcase fits inside the box.
 2. The chocolate fits inside the box.
 3. The box is bigger than the box of chocolates.
 4. The chocolate fits inside the suitcase.
- Q: Does the box fit in the chocolate?
A: no S: 1 4

Task	run1	run2	run3	run4	run5	run6	run7	run8	run9	run10	Mean±std
1:Single Supporting Fact	1.45	0	0	0	0	0	0	0	0	0	0.15±0.44
2:Two Supporting Facts	1.18	1.11	0.68	0.1	0	0.49	1.52	0.71	1.45	1.14	0.84±0.50
3:Three Supporting Facts	1.41	1.42	1.75	0.01	1.48	1.11	0	1.49	1.96	2.12	1.28±0.69
4:Two Arg. Relations	1.09	0.87	2.75	0	0	0.41	0	0	1.48	0	0.66±0.87
5:Three Arg. Relations	1.54	0.46	1.52	1.73	1.54	0.84	0	1.71	1.83	0.47	1.16±0.63
6:Yes/No Questions	0	0	0	0	0	0	0	0	0	0	0.00±0.00
7:Counting	1.04	0.96	0	1.69	0.46	1.34	1.42	1.86	1.65	1.04	1.15±0.55
8:Lists/Sets	2.78	1.89	2.21	1.23	1.29	0.87	1.74	0.94	1.86	0.51	1.53±0.65
9:Simple Negation	0	0	0	0	0	0	0	0	0.09	0	0.01±0.03
10:Indefinite Knowledge	2.18	1.28	0.31	0.1	0	0	0	0	0.65	0.24	0.48±0.69
11:Basic Coreference	0	0	0	0	0	0	0	0	0	0	0.00±0.00
12:Conjunction	0	0	0	0	0	0	0	0	0	0	0.00±0.00
13:Compound Coref.	0	0	0	0	0	0	0	0	0	0	0.00±0.00
14:Time Reasoning	1.89	1.48	2.09	0	2.42	0	0	0	0	0	0.79±0.99
15:Basic Deduction	0	0	0	0	1.87	0	0	0	1.65	2.54	0.61±0.95
16:Basic Induction	52.34	50.45	46.96	50.91	53.44	50.64	52.35	52.71	53.97	50.66	51.44±1.90
17:Positional Reasoning	1.91	0	0	2.51	2.68	1.87	1.95	1.74	0.77	1.68	1.51±0.90
18:Size Reasoning	0.42	0	0	0.12	0.21	0.61	0	1.71	0.84	0.08	0.40±0.51
19:Path Finding	1.48	0	0.87	0.58	0	0.86	0.15	1.4	0.79	0	0.61±0.54
20:Agent’s Motivations	1.69	0	0	0	0	0	0	0	1.56	1.06	0.43±0.67
Average	3.62	3.00	2.96	2.95	3.27	2.95	2.96	3.21	3.53	3.08	3.15±0.24
Failed task(>5%)	1	1	1	1	1	1	1	1	1	1	

Table 5: Results from 10 test runs of the SDMTR model on 20 bAbI tasks, each consisting of 1k test questions, after 500 epochs of joint training with different seed values. The best run is marked in bold.

E.2 Sort-of-CLEVR

Displayed below is a 75×75 image from the Sort-of-CLEVR dataset, which includes unary, binary and ternary visual question-answering tasks, as shown in Figure 8.

E.3 Detecting Equilateral Triangles

An example of equilateral triangle detection is shown in Figure 9.

Ternary questions
 Q: How many objects are in the rectangle formed by the centers of the red and yellow objects?
 Answer: 1
 Q: Are there any objects on the line formed by the centers of the red and grey objects?
 Answer: yes
 Q: How many objects form an obtuse triangle with a blue object and a yellow object?
 Answer: 3

Binary questions:
 Q: What is the shape of the object that is furthest from the green object?
 Answer: orange
 Q: What is the color of the object that is closest to the red object?
 Answer: green
 Q: How many objects have the shape of the yellow object?
 Answer: 3

Unary questions:
 Q: What is the shape of the red object?
 Answer: circle
 Q: Is the green object on the left or right of the image?
 Answer: left
 Q: Is the grey object on the top or bottom of the image?
 Answer: top

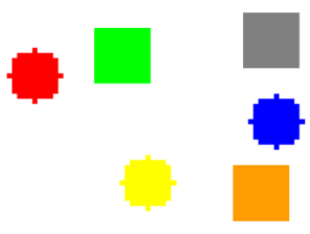


Figure 8: An instance from the Sort-of-CLEVR dataset.

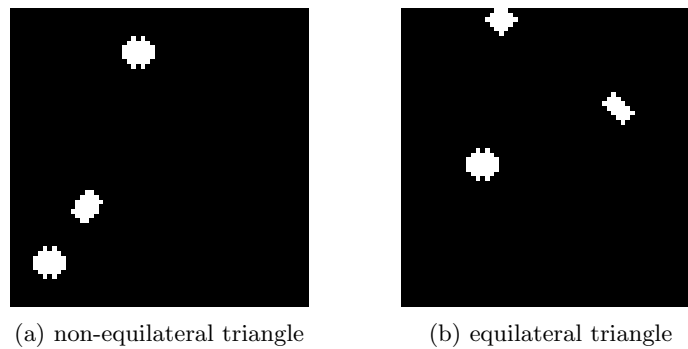


Figure 9: An illustration of the equilateral triangle detection task.