Delayed Attention Training Improves Length Generalization in Transformer–RNN Hybrids

A research paper discussion

Paper authors: Buu Phan, Reza Ebrahimi,
Sanjay Haresh, and Roland Memisevic

arXiv:2510.00258v1 | Reading: 13 October 2026

Train the recurrent path first; then enable attention.

Can one model preserve two complementary skills?

State tracking

Update a running state after every new value.

Modulo addition has a compact recursive solution.

LSTM bias: process the sequence step by step.

Associative recall

Retrieve a past value when its key is queried.

Longer sequences contain more associations to distinguish.

Attention bias: access past positions by content.

The question: does joint training produce this division of labor?

Paper, Sec. 1, PDF p. 1.

Each example requires both answers to be correct

$$\langle \text{bos} \rangle \; k_1 v_1 \cdots k_n v_n \; \langle \text{modulo} \rangle \; m \; \langle \text{recall} \rangle \; k_j v_j$$

Teaching example

Prefix: `A 7  B 8  C 4` Query key: `B`

$$m = (7 + 8 + 4) \bmod 10 = \mathbf{9}, \qquad v_j = \mathbf{8}.$$

$$\text{Acc}_{\text{joint}} = \frac{1}{N_{\text{eval}}} \sum_{r=1}^{N_{\text{eval}}} \mathbf{1}[\hat{m}_r = m_r \;\wedge\; \hat{v}_{j,r} = v_{j,r}]$$

One failure makes the whole example incorrect.

Paper, Sec. 2, PDF pp. 1–2. Example and metric notation are explanatory.

A reusable recurrence explains the desired extrapolation

$$s_0 = 0, \qquad s_i = (s_{i-1} + v_i) \bmod 10, \qquad m = s_n$$

$$0 \xrightarrow{+7} 7 \xrightarrow{+8} 5 \xrightarrow{+4} 9$$

- The state has only ten possible values.
- The same update can be repeated on a longer sequence.
- Recall additionally requires the association $\texttt{B} \mapsto 8$.

Why a hybrid is attractive

A compact running state and access to past associations can coexist.

Teaching explanation of the task in Sec. 2; not a measured hidden-state trajectory.

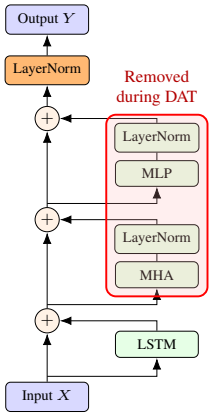
Five hybrids combine the same component types

Architecture	Computation / flow
$\mathrm{Attn} \circ \mathrm{LSTM}$	$H = \mathrm{LSTM}(X),\ Z = \mathrm{Attn}(H)$
$\mathrm{LSTM} \circ \mathrm{Attn}$	$H = \mathrm{Attn}(X),\ Z = \mathrm{LSTM}(H)$
$\mathrm{LSTM} + \mathrm{Attn}$	$Z = H_{\mathrm{lstm}} + H_{\mathrm{attn}}$
Hybrid Block	LSTM, MHA, MLP within a residual block
$\mathrm{LSTM} \circ \mathrm{Attn} \circ \mathrm{LSTM}$	LSTM $\rightarrow$ attention $\rightarrow$ LSTM

Function composition is read right to left.

Paper, Appendix A, PDF p. 4. Attn includes MHA + MLP. Sec. 2 says four variants; Fig. 1 and Appendix A list five.

DAT removes the attention module's early contribution



- Read the original diagram from bottom to top.
- The LSTM path remains available during the first phase.
- The red region includes MHA, MLP, and their LayerNorms.
- Later, attention and recurrence train jointly.

Forward contribution is disabled during the early phase.

Original crop: Fig. 1 left, PDF p. 2. The caption repeats the red attention module $\times 4$.

The recurrent path receives a 1500-epoch head start

Phase 1	Phase 2a	Phase 2b
Attention contribution off	Enable attention	Full hybrid active
Train LSTM component	Joint training	Continue joint training
1500 epochs Learning rate 10^{-4}	**500 epochs** Learning rate 10^{-4}	**Duration unspecified** Learning rate 10^{-5}

Adam; $2^{14} = 16{,}384$ randomly generated samples per epoch.

Ordinary training: 10^{-4} for 500 epochs, then 10^{-5}.

Paper, Sec. 2, PDF p. 2; Appendix A, PDF p. 4.

The evaluation separates joint success from single skills

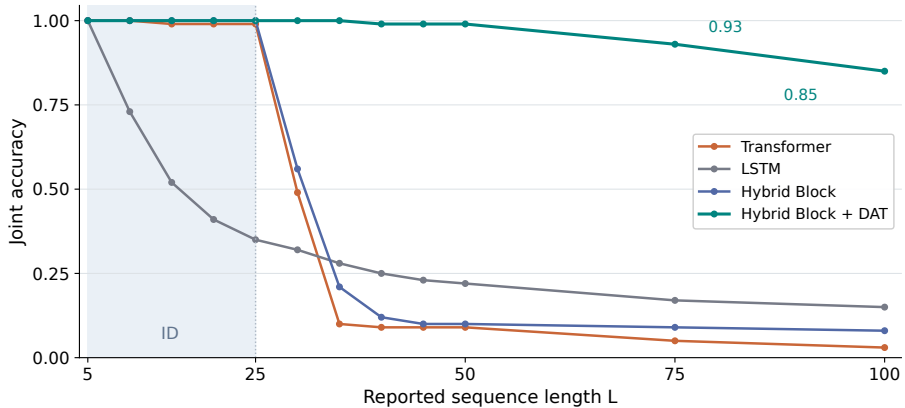
- **Shared model settings:** embedding dimension $d = 256$; default RoPE; Transformer components use four blocks by default.
- **Joint task:** two pure baselines and five hybrids; hybrids compared with and without DAT.
- **Single tasks:** recall-only and modulo-only models are trained and evaluated separately; no DAT rows in these tables.
- **Length:** reported ID range $L \leq 25$; OOD points extend to 100.

Interpret the length scale as reported

The paper does not explicitly map table length L to total token count. The threefold comparison is 75/25.

Paper, Sec. 2 / Fig. 1, PDF p. 2; Appendices A–B, PDF p. 4.

DAT preserves high joint accuracy beyond the ID boundary



Hybrid Block at $L = 75$: 0.09 $\rightarrow$ **0.93** (**+84 percentage points**).

Redrawn from Fig. 1 right, PDF p. 2. Higher is better; shaded region is ID. Lines connect reported points.

Four of five DAT hybrids exceed 90% at threefold length

	$L = 75$		$L = 100$	
Architecture	Ordinary	DAT	Ordinary	DAT
$\mathrm{LSTM} \circ \mathrm{Attn}$	0.05	**0.93**	0.04	**0.88**
$\mathrm{Attn} \circ \mathrm{LSTM}$	0.07	0.80	0.05	0.56
Hybrid Block	0.09	**0.93**	0.08	0.85
$\mathrm{LSTM} \circ \mathrm{Attn} \circ \mathrm{LSTM}$	0.06	**0.91**	0.06	0.80
$\mathrm{LSTM} + \mathrm{Attn}$	0.12	**0.95**	0.08	0.85

Joint accuracy, 0–1. All five DAT hybrids reach 0.99 at $L = 50$.

Paper, Fig. 1 right, PDF p. 2. Architecture still matters at longer lengths.

Recall-only results reveal the LSTM retrieval bottleneck

Model	$L = 25$	$L = 50$	$L = 75$	$L = 100$
LSTM	0.83	0.53	0.37	0.28
Transformer	1.00	0.99	0.90	0.80
Hybrid Block	1.00	0.99	0.98	**0.92**

- Attention-based models retrieve well beyond length 25.
- Their recall also degrades at the longest tested lengths.
- These are separately trained recall-only models.

Selected rows from Table 1, Appendix B, PDF p. 4. Single-task accuracy; no DAT rows.

Adding attention loses the LSTM's modulo extrapolation

Model	$L = 25$	$L = 30$	$L = 50$	$L = 100$
LSTM	1.00	1.00	**1.00**	**1.00**
Transformer	0.98	0.51	0.10	0.10
Hybrid Block	1.00	0.45	0.10	0.10
$\mathrm{Attn} \circ \mathrm{LSTM}$	1.00	0.63	0.10	0.10

The informative failure

The recurrent component remains in the hybrid, but ordinary training no longer yields the observed LSTM extrapolation.

Selected rows from Table 2, Appendix B, PDF p. 4. Modulo-only accuracy; no DAT rows.

The evidence supports a training-path explanation

1. **Capability:** LSTM alone solves the tested modulo lengths.
2. **Interference:** ordinary hybrids fail on longer modulo sequences.
3. **Intervention:** delaying attention strongly improves joint-task OOD accuracy.

Authors' interpretation

Early attention enables shortcut solutions; DAT first establishes a more robust recurrent computation.

Paper, Sec. 2, Fig. 1 and Table 2, PDF pp. 2 and 4. The exact shortcut and internal state representation are not measured.

More candidates dilute attention at a fixed score margin

$N \geq 2$ positions: correct score a, distractor scores b; $\Delta = a - b$.

$$p_\star = \frac{e^a}{e^a + (N-1)e^b} = \frac{1}{1 + (N-1)e^{-\Delta}}$$

- At fixed Δ, the correct position receives less weight as N grows.
- For target weight $0 < \rho < 1$, maintaining $p_\star \geq \rho$ requires $\Delta \geq \log(N-1) + \log\frac{\rho}{1-\rho}$.

How far the paper's evidence goes

The authors attribute the long-length drop to softmax sharpness. Table 1 shows recall degradation; it does not measure sharpness or isolate this cause in the DAT model.

Paper's attribution: Sec. 2, PDF p. 2. Equations are a teaching derivation assuming equal distractor scores.

Training budget and implementation detail limit attribution

- **Budget:** DAT's first phase processes $1500 \times 2^{14} = 24{,}576{,}000$ generated samples. Final epoch counts and matched total-budget controls are unspecified.
- **Reproduction:** the precise loss mask, decoding protocol, and architecture-specific bypass implementations are not provided.
- **Scope:** synthetic LSTM–Transformer tasks at $d = 256$; no language-modeling result or reported seed variability.

Paper, Sec. 2, PDF p. 2; Appendix A, PDF p. 4. Sample count is arithmetic from the reported schedule.

Three experiments would make the mechanism clearer

1. **Match total training budget and vary the delay.**
   Separate scheduling effects from total samples or compute.
2. **Probe the same joint-task model.**
   Measure both subtask accuracies and decode the running remainder.
3. **Intervene on retrieval sharpness.**
   Track score margins and target weights; test a temperature intervention.

Reading discussion: proposed follow-ups, not experiments reported in the paper.

Training order is part of hybrid-model design

Main result

Delaying attention strongly improves length generalization on a task requiring both state tracking and associative recall.

- Final architecture alone does not determine which solution is learned.
- At $L = 75$, four DAT variants score 0.91–0.95; one scores 0.80.
- Farther extrapolation and transfer to language modeling remain open.

Paper, Fig. 1 and Sec. 3, PDF p. 2.

Primary source and reading map

Buu Phan, Reza Ebrahimi, Sanjay Haresh, Roland Memisevic.

Delayed Attention Training Improves Length Generalization in Transformer–RNN Hybrids.

arXiv:2510.00258v1, submitted 30 September 2025.
https://arxiv.org/abs/2510.00258v1

- PDF p. 2: task, DAT, architecture diagram, joint results.
- PDF p. 4: training details, architecture definitions, single tasks.
- Teaching examples and follow-up ideas are identified on their slides.

Backup: DAT across the reported length transition

DAT architecture	25	30	35	50	75	100
$\mathrm{LSTM} \circ \mathrm{Attn}$	1.00	0.99	0.99	0.99	0.93	0.88
$\mathrm{Attn} \circ \mathrm{LSTM}$	1.00	0.99	0.99	0.99	0.80	0.56
Hybrid Block	1.00	1.00	1.00	0.99	0.93	0.85
Sandwich	1.00	1.00	0.99	0.99	0.91	0.80
Parallel	1.00	1.00	1.00	0.99	0.95	0.85

Joint accuracy. Sandwich: $\mathrm{LSTM} \circ \mathrm{Attn} \circ \mathrm{LSTM}$.

Parallel: $\mathrm{LSTM} + \mathrm{Attn}$. Column headers are reported lengths.

Selected columns from Fig. 1 right, PDF p. 2.

Backup: disabling contribution requires a suitable bypass

Serial composition

Preserve the information path when the attention module is absent.

$$A_{\text{early}}(H) = H$$

Parallel addition

Remove the branch contribution before summing.

$$Z_{\text{early}} = H_{\text{lstm}} + 0$$

- These illustrate bypass semantics; the paper does not supply this code.
- Keeping a frozen attention output active changes the forward function.
- A switch to modulo-only supervision is not specified for DAT.

Implementation interpretation of Sec. 2 and Fig. 1, PDF p. 2; conceptual formulas, not author pseudocode.