

AMEMOPS: EVALUATING MEMORY UPDATES IN LONG AUDIO STREAMS

Peiwei Chang¹, Li Zhang², Mark D. Plumbley³, Philip J.B. Jackson¹, Wenwu Wang¹

¹Centre for Vision, Speech and Signal Processing, University of Surrey, UK

²Department of Computer Science, Royal Holloway, University of London, UK

³Department of Informatics, King’s College London, UK

ABSTRACT

Audio memory must recognise recurring sources across acoustic changes while keeping their attributes up to date. Overall accuracy does not reveal how these demands interact. We introduce Audio Memory Operations (AMemOps), a benchmark for latest-attribute retrieval, preservation of unchanged attributes and absent-source rejection in long audio streams. Paired controls vary recordings, training identities and earlier events to examine their effects on updating. Training on other recordings of the test instruments increases a reference delta-rule memory’s update advantage over matched attention by 13.4 percentage points. At retrieval delays of 120 s, matched histories separate benefits of consistent repetition from costs of conflicting earlier values, while additive-memory controls identify benefits of erasure.

Index Terms— associative memory, linear attention, delta rule, audio understanding, memory updates

1. INTRODUCTION

Understanding long audio streams requires recognising recurring sources and tracking their changing attributes [1]. When a violin returns at a different pitch and level, a system should retrieve its current level while retaining its unchanged spatial position. Such tracking could support persistent audio assistants and acoustic monitoring, where new observations can make earlier descriptions outdated. Retrieval errors can therefore arise from failures to recognise the source, access its history or update an attribute.

Existing models offer different ways to retain a source’s history. Attention stores previous observations for later retrieval, whereas delta-rule memories update key–value associations in a fixed-size state [2, 3]. Recent variants refine how this state is erased [4, 5]. Both depend on reliable representations: attention must select the latest matching occurrence, while recurrent updates require consistent source keys [6, 7].

Existing benchmarks address parts of this problem. Explicit-key tasks test retention and replacement [8, 9, 10]; conversational MemOps tests whether answers reflect updated facts [11]. Neither requires recognising a source across acoustic recordings. Audio benchmarks cover attributes, timing and long-context retrieval [12, 13, 14, 15, 16, 17, 18]. EnvMem diagnoses retrieval of an earlier sound [19], while EmoUpdate addresses interference from prior emotion labels supplied as text [20]. These studies leave open how recording variation and obsolete acoustic observations jointly affect retrieval of a source’s latest attributes.

We introduce the **Audio Memory Operations (AMemOps)** framework (Fig. 1). Our main contributions are: **(1) An audio memory benchmark:** environmental-sound and instrument streams test

latest-attribute retrieval, unchanged-attribute retention and absent-source rejection without source labels. **(2) Paired diagnostics:** recording and training controls test acoustic variation and source familiarity, while matched histories vary earlier attributes with the latest event, query and answer fixed. **(3) Controlled model comparisons:** we compare attention and recurrent memory under these conditions, using additive-memory controls and source-specific state edits to examine erasure and selective access.

2. METHODS

Audio streams and queries. Each sound class or instrument identity has independently controlled gain and stereo pan (Fig. 1a). Nonoverlapping excerpts, distractors and queries form 2,048-frame streams at 25 Hz (81.92 s). Recordings are normalised to root mean square (RMS) amplitude 0.05 before extracting 18–32-frame excerpts (0.72–1.28 s). Excerpts receive 10-ms fades, uniformly sampled gain $\{-18, -12, -6, 0\}$ dB and equal-power pan $\{0, 1/3, 2/3, 1\}$ (left to right). Panning preserves total event energy. We then add independent white Gaussian noise per channel, 45 dB below the RMS reference, without final normalisation or clipping.

A 12-frame query requests the source’s latest gain or pan, or *none* for a previously absent source. Queries use -3 dB and centre pan, outside the answer classes, to avoid revealing the answer. Query frames update model memory but not ground-truth attributes. Inputs combine stereo log-mel features with query, gain-request and pan-request flags. Features use 64 bands per channel (16 kHz, 1,024-sample Hann window, 640-sample hop, 50–7,800 Hz). Separate gain and pan classifiers predict four values plus *none*, using cross-entropy at query end.

The generator proposes 10 initial-value (no prior change), 8 post-update, 6 similar-source interference and 2 absent-source queries per stream, dropping events that do not fit. Post-update queries request gain and pan separately; other queries select either uniformly. A retained attribute has never changed, although the other has. Proposed onset offsets between related events are log-uniform between 2 and 60 s. Retrieval delay runs from the latest source event’s end to query end.

Memory models. Gated DeltaNet-2 (GDN2) [23] combines key-directed erasure [25] and recurrent linear attention [26]. Separate decay, erase and write gates enable controlled ablations.

Each head updates $S_t \in \mathbb{R}^{d_k \times d_v}$ from learned query and key vectors $q_t, k_t \in \mathbb{R}^{d_k}$ and value $v_t \in \mathbb{R}^{d_v}$:

$$S_t = (I_{d_k} - k_t(b_t \odot k_t)^\top) \text{Diag}(\alpha_t) S_{t-1} + k_t(w_t \odot v_t)^\top, \quad (1)$$

$$o_t = S_t^\top q_t. \quad (2)$$

Decay α_t scales old contents, b_t erases along the current key, and w_t writes the new key–value association (Fig. 1b). Readout returns

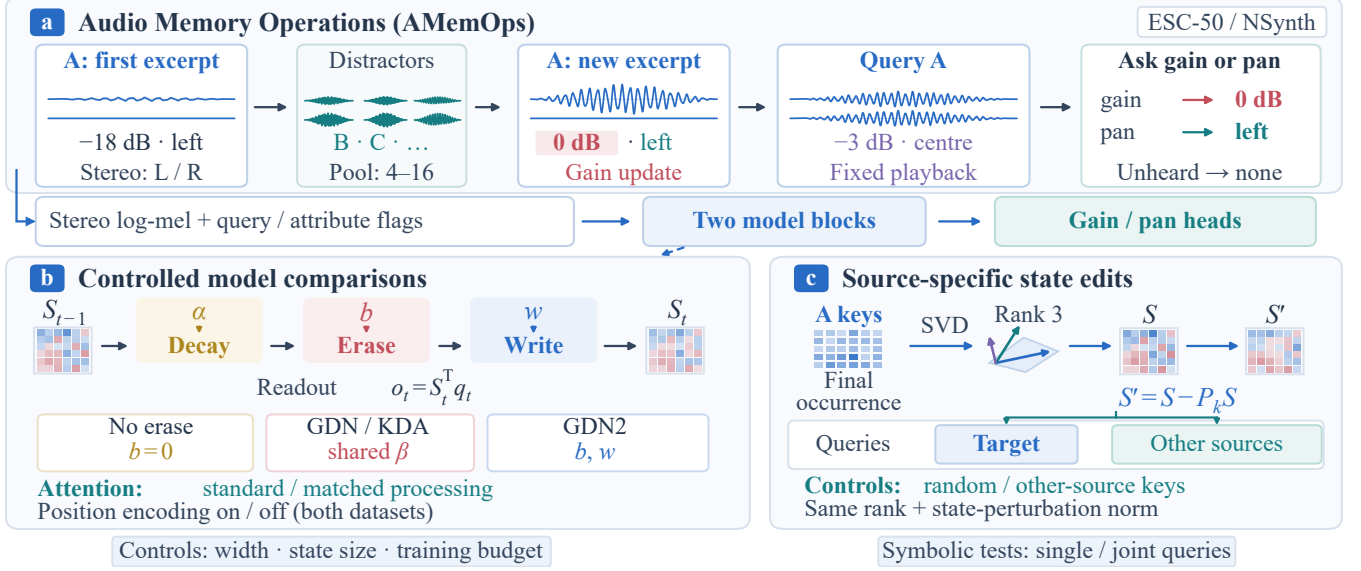


Fig. 1. AMemOps overview. (a) Latest gain or pan queries; source labels are illustrative, not inputs. (b) Decay, erasure, writing and readout, comparing Gated DeltaNet (GDN) [21], Kimi Delta Attention (KDA) [22], Gated DeltaNet-2 (GDN2) [23] and attention. (c) Source-key state edits test selective access against norm-matched controls.

Table 1. Model components: two pre-normalised blocks, width $d=128$. Conv-4 is four-frame depthwise causal convolution, and output normalisation is per head. SiLU: sigmoid linear unit; RMSNorm: root mean square layer normalisation; RoPE: rotary position embeddings [24].

Mixer component	Attention		Recurrent
	Standard	Matched	
Heads $\times$ dimension	4×32	2×64	2×64
Q/K/V processing	None	Conv-4 + SiLU	Conv-4 + SiLU
Q–K normalisation	None	L_2	L_2
Output norm. and gate	None	RMSNorm + SiLU	RMSNorm + SiLU
RoPE	Optional	Optional	None

$o_t \in \mathbb{R}^{d_v}$ from the updated state. Gates satisfy $\alpha_t, b_t \in [0, 1]^{d_k}$ and $w_t \in [0, 1]^{d_v}$. I , $\odot$ and Diag denote identity, elementwise product and diagonal-matrix construction.

Kimi Delta Attention (KDA) [22] ties $b_t = \beta_t \mathbf{1}_{d_k}$ and $w_t = \beta_t \mathbf{1}_{d_v}$ with scalar β_t . Scalar decay yields Gated DeltaNet (GDN) [21]. *No-erase* sets $b_t=0$ with channel-wise writing and decay [27]. *Mamba-2-style* also sets $b_t=0$, fixes $w_t = \mathbf{1}_{d_v}$ and uses scalar decay [28]. *Matched attention* uses causal softmax with the recurrent models’ head dimensions and input/output processing (Table 1).

Recording controls. Fixed recordings replay each source’s first recording from the start in events and queries, preserving timing, attributes, answers and noise. Intermediate controls cyclically reuse two or four recordings per source; an unrestricted control uses all recordings, also starting at zero.

Matched event histories. We vary the earlier target event while fixing the latest event, query and answer. The earlier target matches both latest attributes (no update), or differs only in the queried (updated) or other attribute (retained). We evaluate the 24k models in Table 2 on NSynth instrument identities seen during training. All operation, delay, load and recording conditions share 192 single-query templates: two repeats per queried attribute, latest gain–pan pair and

Table 2. Query-weighted accuracy (%) on 48 AMemOps streams after 12,000 (12k) and 24,000 (24k) training steps, averaged over three seeds.

model	all queries		query type, 24k	
	12k	24k	absent	updated
<i>ESC-50 environmental sounds</i>				
GDN2	68.8	69.1	60.1	67.5
KDA	67.3	67.8	58.0	65.8
GDN	65.3	67.0	49.7	65.4
no-erase	63.6	65.3	49.3	63.2
Mamba-2-style	61.4	65.5	50.7	60.1
standard + RoPE	66.2	67.9	43.4	65.1
standard, no RoPE	—	66.9	38.9	62.9
matched + RoPE	—	71.9	55.6	69.6
matched, no RoPE	—	72.7	59.0	67.0
<i>NSynth instruments</i>				
GDN2	76.2	90.4	85.4	88.8
KDA	73.3	90.3	84.4	88.7
GDN	77.7	93.6	90.3	92.8
no-erase	67.9	85.9	74.0	81.6
Mamba-2-style	64.5	78.1	61.5	68.4
standard + RoPE	56.3	68.8	53.5	61.7
standard, no RoPE	—	62.0	41.7	53.2
matched + RoPE	—	87.1	77.4	73.6
matched, no RoPE	—	86.4	80.6	69.5

possible value change.

Two 0.96-s target events start 4 s apart, the latest ending at 20 s. Queries end at 40 or 140 s in 4,096-frame inputs (163.84 s): delays of 20 or 120 s, within or beyond the training context.

Sixteen distinct distractor sources appear once at evenly spaced onsets between target and query. Low load retains events 1, 6, 11 and 16. Longer delays stretch spacing, keeping counts, identities and durations fixed. Noise is shared. At 120 s, original-recording

Table 3. Mean per-stream NSynth accuracy (%) on original (O) and fixed (F) recordings: 192 held-out streams per set, 24k models from Table 2, three seeds.

Model	Initial		Updated		Retained	
	O	F	O	F	O	F
<i>Seen identities (Set 1)</i>						
GDN2	92.2	99.4	89.2	98.8	94.1	99.5
GDN	93.6	98.7	92.0	98.8	95.6	99.2
no-erase	88.9	98.6	80.9	95.3	92.7	99.1
matched + RoPE	91.8	99.6	77.1	98.7	94.9	99.7
matched, no RoPE	91.8	99.7	72.2	98.1	94.5	99.4
<i>Unseen identities (Set 2)</i>						
GDN2	43.9	88.4	39.4	87.6	53.3	92.6
GDN	35.4	69.9	34.3	72.8	43.5	80.3
no-erase	44.4	82.3	40.2	76.2	54.8	90.1
matched + RoPE	45.9	94.6	40.9	92.6	58.5	95.3
matched, no RoPE	43.2	93.9	39.2	92.1	56.6	94.9

controls for GDN2, no-erase and no-RoPE attention omit the earlier target or replay the first distractor there, preserving later audio.

State interventions. Source-key projections test selective access (Fig. 1c). Per-head P_k projects onto the top three right singular vectors of target keys at the final occurrence. We apply $S \leftarrow S - P_k S$ in both layers. Eligible queries concern target or other-source bindings established before the edit and not refreshed before querying. Controls use three random subspaces or the most overlapping other-source key subspace. Per-layer rescaling over both heads matches the target edit’s Frobenius norm (relative error $< 10^{-6}$).

3. EXPERIMENTS

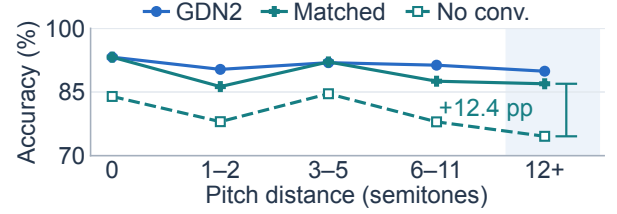
Data. ESC-50 [29] supplies 50 classes: folds 1–4 for training and fold 5 for evaluation. The NSynth models in Tables 2–3 train on official validation notes and evaluate on test notes [30]. Set 1 comprises 53 instruments shared across these splits at MIDI pitches 36–84, with disjoint note IDs (8,096 training, 2,632 evaluation). The generator varies ESC-50 excerpt starts and NSynth recordings. Mean source spectra pooled across training and evaluation recordings guide stream generation: similar pairs use the upper correlation quartile, dissimilar pairs the lower.

Architecture and training. Two recurrent layers each contain two 64×64 states, totalling 64 KiB at 32-bit precision. Matched attention’s learned query–key scales [31] start at $\sqrt{d_k} = 8$. Log-spaced decay rates $-\log \alpha \in [10^{-3}, 10^{-1}]$ initialise multiple timescales; erase and write gates start at $\text{sigmoid}(1)$. AdamW uses batch size 40, peak learning rate 10^{-3} , 300-step warm-up and cosine decay. We report final 12k and 24k checkpoints from three seeds. Training varies recordings, similarity and distractor counts (4, 8 or 16). Attention has 450k parameters; GDN2 has 549k. Symbolic controls use unit-vector identities, two 16-value attributes plus *none*, one layer, one head, two seeds and 20k steps.

Identity controls. We test frozen models on two disjoint sets of 53 unseen instruments (Sets 2 and 3), matched by family, source type and pitch–velocity coverage. Each has 2,632 notes and 192 streams. Calibration recordings with disjoint pitches define similarity pools.

Training familiarity. We train GDN2 and matched RoPE attention separately on Sets 1 and 2, then test on both, keeping each set’s 192 test streams fixed. An instrument is *seen* when training includes other recordings of it, otherwise *unseen*.

(a) Acoustic processing



(b) Seen vs unseen in training

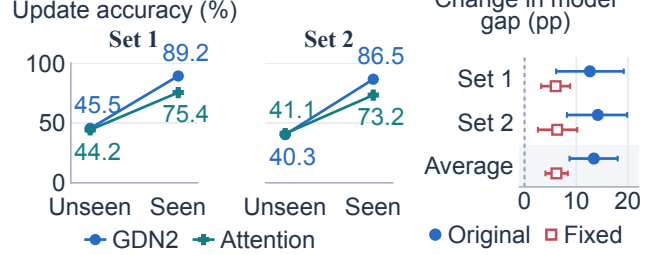


Fig. 2. NSynth (24k, three seeds). (a) Pitch distance from query to latest same-source event. Dashed: matched attention without convolution (9–48 streams/bin, at least two queries/stream). (b) Left: update accuracy on identical test streams for identities seen/unseen in training. Right: change in the GDN2–attention gap (Eq. 3). Attention uses RoPE; bars show 95% CIs.

These models (Fig. 2b) use 6,454 training, 1,589 calibration and 2,632 evaluation recordings per set, excluding shared note IDs and exact training–evaluation waveform duplicates. Calibration recordings define similarity pools. Training uses 24k steps and three seeds per training set (12 runs).

Let $g_{c,s}$ be the GDN2–attention update accuracy gap on test set c under training status s . Its mean change with familiarity is:

$$I = \frac{1}{2} \sum_{c=1}^2 (g_{c,\text{seen}} - g_{c,\text{unseen}}). \quad (3)$$

Statistics. Paired analyses average eligible gain and pan queries within streams, then weight non-empty streams and seeds equally. Attribute-specific scores repeat this averaging on their own queries. Differences are in percentage points (pp), with 95% confidence intervals (CIs). In 100,000 bootstrap resamples, seeds are drawn independently across models and training sets; streams are paired across models and conditions but independent across test sets. Four primary history contrasts test delay and load effects on GDN2’s update gaps against no-erase and no-RoPE attention, averaging over the other factor on original recordings. Simultaneous intervals use the 95th percentile of the maximum absolute centred bootstrap error.

3.1. Acoustic processing and recording variation

Input processing. GDN leads on instruments, while matched attention leads on environmental sounds (Table 2). Matching recurrent heads and processing improves NSynth attention accuracy by 21.4 pp [19.0, 23.7], averaged over RoPE settings. Convolution helps retrieval across large pitch differences (Fig. 2a).

Recording reuse improves retrieval. Fixed recordings improve all models and close update gaps on seen instruments (Table 3). With original recordings, GDN2 leads matched attention on updates by 12.1 and 17.0 pp with and without RoPE. Retained-attribute differences are -0.8 and -0.4 pp (CIs include zero). Fixing recordings

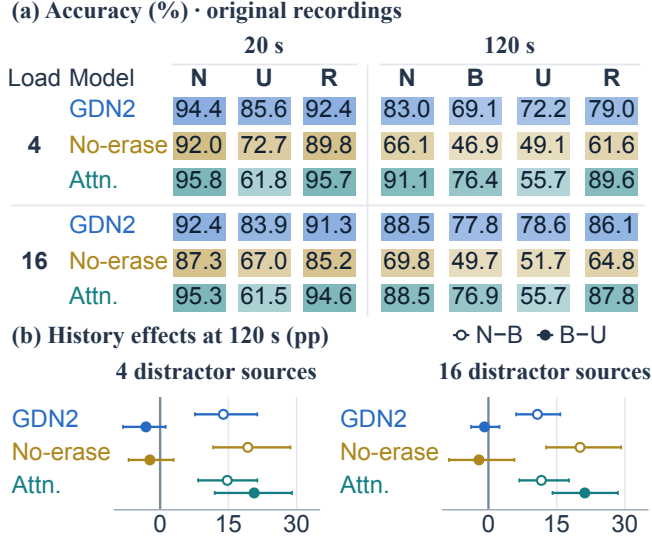


Fig. 3. NSynth Set 1, Table 2 models at 24k, three seeds. (a) Accuracy: consistent (N), earlier target absent (B), updated (U), retained (R). Load counts distractor sources. (b) Repetition benefit N–B and net conflict cost B–U, with marginal 95% CIs. Each condition has 192 paired queries. Attention has no RoPE.

improves matched RoPE attention’s gain and pan update accuracy by 20.6 and 20.3 pp, respectively. With two or four recordings per source, the gap remains 12–13 pp; with one, both models approach perfect accuracy. On original recordings, excluding queries whose own or earlier same-source recordings match training audio (same instrument and pitch; first-second waveform cosine > 0.9999) retains 12.0 and 16.1 pp advantages (616 updates, 185 streams).

3.2. Seen and unseen instruments

On unseen Set 2 instruments, GDN2’s update gaps against matched attention are -1.5 pp with RoPE and $+0.2$ pp without it (CIs include zero). The gap against RoPE attention is $+1.8$ pp $[-2.3, 5.9]$ on Set 3. Initial and retained accuracy also fall (Table 3), showing difficulty beyond updates. Recording reuse improves RoPE attention’s pan update accuracy by 50.9–54.1 pp across the unseen sets.

Training on other recordings of the same test instruments increases GDN2’s update advantage over matched RoPE attention by $I = 13.4$ pp $[8.7, 18.0]$ (Fig. 2b): 9.6 pp $[3.8, 15.3]$ for gain and 16.8 pp $[11.3, 22.4]$ for pan. Fixed recordings reduce I to 6.2 pp $[4.0, 8.4]$, a paired reduction of 7.2 pp $[3.1, 11.4]$.

3.3. Conflicting history at long retrieval delays

At 120 s with four distractors, matched no-RoPE attention leads GDN2 without updates but trails after updating (Fig. 3). Across delays and loads, its update loss relative to no update is 33–35 pp, versus 9–11 pp for GDN2. Retained-attribute losses are only 0–5 pp.

At 120 s, attention accuracy without the earlier target is 76.4% with four distractors and 76.9% with sixteen.

Against these baselines, conflicting history reduces accuracy by 20.7 pp $[12.0, 29.0]$ and 21.2 pp $[14.1, 28.5]$, while consistent history adds 14.8 and 11.6 pp. Replacing the earlier target with a distractor gives similar costs (20.1 and 21.0 pp). GDN2 and no-erase show no clear net loss (CIs include zero).

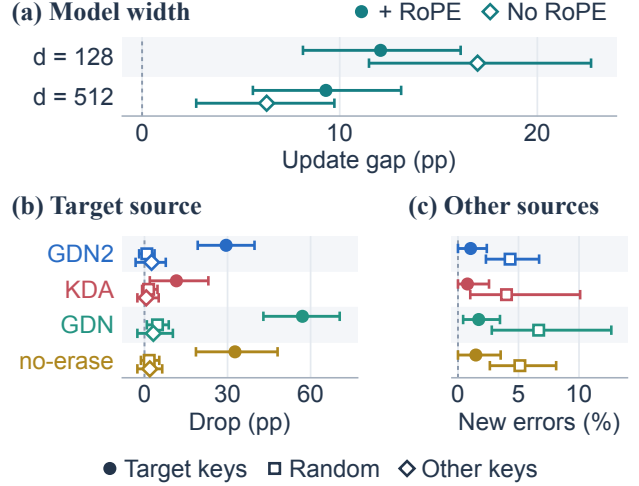


Fig. 4. (a) GDN2’s update advantage over matched attention (NSynth, 24k, 929 queries in 192 streams). (b,c) Rank-3 edits (ESC-50, 12k): 52 target and 148 other-source queries in 26 streams before target reappearance. (b) Net target accuracy loss. (c) Correct-to-wrong transitions divided by all eligible other-source queries. Stream-weighted means, three seeds, 95% CIs.

Previous values account for 78% of no-RoPE attention’s update errors and 23% of GDN2’s at 120 s with four distractors (pooled seeds). The four pre-specified interactions show no clear increase in GDN2’s update advantage with delay or load. RoPE attention falls below 3% on all operations at 120 s.

3.4. Erasure, capacity and selective access

Erasure improves updating. With fixed recordings, GDN2 exceeds no-erase on updates by 11.3 pp $[1.9, 24.8]$ on Set 2 and 9.5 pp $[0.3, 20.4]$ on Set 3. Excluding Set 3 streams with duplicate evaluation notes leaves a 9.0 pp gap $[-0.2, 19.9]$ (815 updates in 172 streams). The advantage is 5.5 pp $[3.4, 7.7]$ on 192 ESC-50 streams with repeated updates (10 update sources, up to four updates each). Symbolic single and joint queries likewise yield 82.7–85.1% per-attribute update accuracy for GDN2, KDA and GDN, versus 58.2% for no-erase.

Width control. Widening d from 128 to 512 with the same two 64-dimensional heads preserves GDN2’s update advantage: 9.3 pp $[5.6, 13.1]$ with RoPE and 6.3 pp $[2.7, 9.7]$ without it (Fig. 4a).

Selective access. Source-key edits cause larger target losses and fewer other-source errors than random edits in erasing and additive memories (Fig. 4b,c).

4. CONCLUSION

AMemOps evaluates retrieval of current attributes and preservation of unchanged ones. Its paired controls show how acoustic variation and familiarity shape model comparisons, and distinguish benefits of consistent repetition from costs of conflicting observations. Additive controls show when erasure improves updating.

5. ACKNOWLEDGEMENTS

This work was supported by the UKRI AI Centre for Doctoral Training in AI for Digital Media Inclusion [grant number EP/Y030915/1] and by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/Y028805/1]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

6. REFERENCES

- [1] B. G. Shinn-Cunningham, “Object-based auditory and visual attention,” *Trends Cogn. Sci.*, vol. 12, no. 5, pp. 182–186, 2008.
- [2] I. Schlag, K. Irie, and J. Schmidhuber, “Linear transformers are secretly fast weight programmers,” in *Proc. ICML*, 2021, vol. 139, pp. 9355–9366.
- [3] S. Yang, B. Wang, Y. Zhang, Y. Shen, and Y. Kim, “Parallelizing linear transformers with the delta rule over sequence length,” in *Proc. NeurIPS*, 2024, vol. 37.
- [4] X. Li, C. Zhang, H. Luo, X. Lin, Z. Wang, Z. Qiu, et al., “Erase-then-Delta attention: Decoupling erase and write addresses in delta-rule linear attention,” *arXiv:2606.26560*, 2026.
- [5] D. Gupta, A. Das, and D. Gupta, “The query knows what to forget: A second erase direction for linear attention,” *arXiv:2608.13668*, 2026.
- [6] A. Arora, N. Rathi, N. R. Selvam, R. Csordás, D. Jurafsky, and C. Potts, “Mechanistic evaluation of transformers and state space models,” *arXiv:2505.15105*, 2025.
- [7] A. Ghriess and S. Chakraborty, “Kernelized linear attention: Breaking the capacity wall with symmetric cones,” *arXiv:2607.17419*, 2026.
- [8] S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, et al., “Zoology: Measuring and improving recall in efficient language models,” in *Proc. ICLR*, 2024.
- [9] M. Poli, A. W. Thomas, E. Nguyen, P. Ponnusamy, B. Deiseroth, K. Kersting, et al., “Mechanistic design and scaling of hybrid architectures,” in *Proc. ICML*, 2024, vol. 235, pp. 40908–40950.
- [10] M. Dowling, H. Jeon, C. Savin, and I. M. Park, “Memory by design: Probabilistic sequence layers,” *arXiv:2605.31163*, 2026.
- [11] X. Hao, Z. Zhang, Z. Lin, Y. Sun, Z. Guo, X. Zhang, et al., “MemOps: Benchmarking lifecycle memory operations in long-horizon conversations,” *arXiv:2607.12893*, 2026.
- [12] H. M. Fayek and J. Johnson, “Temporal reasoning via audio question answering,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2283–2294, 2020.
- [13] J. Abdelnour, J. Rouat, and G. Salvi, “NAAQA: A neural architecture for acoustic question answering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4997–5009, 2023.
- [14] H. Jin, S. Wang, Y. Xiao, J. Dong, S. Tan, K. Peng, et al., “RAIL: Rethinking auditory intelligence in large audio-language models with a CHC-grounded benchmark,” *arXiv:2606.11260*, 2026.
- [15] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, et al., “Audio Flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities,” in *Proc. ICML*, 2025, vol. 267, pp. 19358–19405.
- [16] P. He, Z. Wen, Y. Wang, Y. Wang, X. Liu, J. Huang, et al., “AudioMarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio LLMs,” *arXiv:2510.07293*, 2025.
- [17] K. Hegde, A. K. Sridhar, N. Vakada, Y. Guo, and E. Visser, “Event-grounded question answering over long audio via structured retrieval,” *arXiv:2602.14612*, 2026.
- [18] A. Gosai, T. Vuong, U. Tyagi, S. Li, W. You, M. Bavare, et al., “Audio MultiChallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction,” in *Proc. ACL*, 2026, pp. 35740–35770.
- [19] Y. Xiao, S. Wang, H. Yin, H. Jia, V. Sethu, E.-J. Holden, et al., “Why can’t they remember? uncovering representation and retrieval bottlenecks in multi-turn acoustic memory,” *arXiv:2605.27039*, 2026.
- [20] H. Liu, Z. Wang, Y. Chen, H. Deng, and X. Tang, “Do SpeechLMs hear their own opinions? diagnosing and mitigating previous-belief contamination in streaming emotion understanding,” *arXiv:2608.20769*, 2026.
- [21] S. Yang, J. Kautz, and A. Hatamizadeh, “Gated delta networks: Improving Mamba2 with delta rule,” in *Proc. ICLR*, 2025.
- [22] Kimi Team, “Kimi Linear: An expressive, efficient attention architecture,” *arXiv:2510.26692*, 2025.
- [23] A. Hatamizadeh, Y. Choi, and J. Kautz, “Gated DeltaNet-2: Decoupling erase and write in linear attention,” *arXiv:2605.22791*, 2026.
- [24] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, “RoFormer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, pp. 127063, 2024.
- [25] A. Graves, G. Wayne, and I. Danihelka, “Neural Turing machines,” *arXiv:1410.5401*, 2014.
- [26] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers are RNNs: Fast autoregressive transformers with linear attention,” in *Proc. ICML*, 2020, vol. 119, pp. 5156–5165.
- [27] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim, “Gated linear attention transformers with hardware-efficient training,” in *Proc. ICML*, 2024, vol. 235, pp. 56501–56523.
- [28] T. Dao and A. Gu, “Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality,” in *Proc. ICML*, 2024, vol. 235, pp. 10041–10071.
- [29] K. J. Piczak, “ESC: Dataset for environmental sound classification,” in *Proc. ACM Multimedia*, 2015, pp. 1015–1018.
- [30] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, et al., “Neural audio synthesis of musical notes with WaveNet autoencoders,” in *Proc. ICML*, 2017, vol. 70, pp. 1068–1077.
- [31] A. Henry, P. R. Dachapally, S. S. Pawar, and Y. Chen, “Query-key normalization for transformers,” in *Findings of EMNLP*, 2020, pp. 4246–4253.