MARS: MULTI-RATE AGGREGATION OF RECENCY SIGNALS FOR SEQUENTIAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential recommendation requires both encoding the content of a user's history and deciding how that content should contribute to the next prediction. We study temporal readout as a modelling choice that can be controlled separately from sequence encoding. We introduce **MARS** (**M**ulti-rate **A**ggregation of **R**ecency **S**ignals), a post-encoder operator that uses observed timestamps to construct K recency-weighted summaries. User-conditioned effective decay rates determine the temporal concentration of each head, and a context-dependent gate fuses the summaries with the encoder's final state. The same readout interface supports Transformer and Mamba backbones without changing downstream item scoring. Our analysis characterises a restricted kernel-approximation family and establishes how effective rates distinguish attention profiles and control their expected lag. In the reported results on four sparse benchmarks, the Transformer instantiation improves HR@10 by an average of 19.7% over the strongest content-only Transformer baseline on each dataset. On dense ML-1M, the Mamba instantiation exceeds SIGMA's HR@10 by 3.2%. Matched-backbone ablations show that gains depend on the encoder and dataset, while single-head variants remain competitive in several settings. These findings support explicit temporal readout as a modular means of introducing recency bias, with its value determined jointly by the encoded representation and the interaction history. Code is available at https://anonymous.4open.science/r/MARSS.

1 INTRODUCTION

Sequential recommendation predicts a user's next interaction from an ordered history. Item order captures transitions, but elapsed time changes how those transitions should be interpreted. Two histories with the same item order may span an hour or a year: their positional structure is identical, yet the relevance of recent and distant events can differ. This distinction raises a question that extends beyond the choice of sequence encoder: how should temporal information shape the final representation used for ranking?

Transformers learn content interactions through self-attention (Kang & McAuley, 2018; Sun et al., 2019; Du et al., 2023), while selective state-space models maintain input-dependent memory dynamics (Gu & Dao, 2023; Liu et al., 2024a; 2025). Both support multiple dependencies and time-scales. Time-aware methods additionally incorporate relative intervals, diverse temporal embeddings, or contextual temporal kernels (Li et al., 2020; Cho et al., 2020). These advances make temporal modelling part of sequence representation learning. A complementary choice remains at the output: which encoded positions should contribute to the user vector, and how should their contributions depend on observed elapsed time?

We formulate this choice as *temporal readout*. The encoder determines what contextual information each hidden state contains; the readout determines how those states are combined. An explicit timestamp-conditioned readout exposes a direct control over recency while preserving the encoder's content-mixing mechanism. This separation also makes the interaction between encoding and aggregation testable: the same readout interface can be paired with different encoder families and compared with their corresponding backbone-only representations. Its potential value lies in a reusable temporal inductive bias, rather than a claim that existing encoders cannot represent multiple time-scales.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We introduce **MARS** (**M**ulti-rate **A**ggregation of **R**ecency **S**ignals) to implement this temporal readout. The operator constructs K normalised decay heads over log-compressed elapsed times. A user-conditioned modulation adjusts each head's effective rate, allowing its attention profile to shift with the encoded history. A context-dependent gate combines the resulting summaries, and a residual path retains the encoder's last-position representation. Pairwise Jensen–Shannon and batch-level balance penalties encourage complementary profiles and distributed gate usage. Together, these components provide an explicit parameterisation of temporal aggregation without changing the downstream item-scoring interface. We study Transformer (MARS-T) and Mamba (MARS-M) instantiations on five benchmarks. The reported results favour MARS-T on the four short-history datasets and MARS-M on dense ML-1M, a pattern summarised by a simple dataset-level selection heuristic. Matched-backbone ablations reveal a more specific conclusion: adding the readout helps in several configurations, but the benefit of multiple heads and individual components varies across settings. This evidence motivates choosing temporal aggregation jointly with the encoder. The accompanying analysis explains the readout's admissible kernel family and the effect of its effective rates, providing an operator-level account of the design. Our **contributions** are as follows:

- **A modular temporal readout.** We introduce a post-encoder aggregation operator with user-conditioned recency heads and adaptive fusion, exposing temporal weighting through a common interface for Transformer and Mamba representations.
- **An analytical characterisation.** We establish finite-mixture approximation for a specified class of Laplace-mixture kernels and characterise the relationship between effective rates, distinct attention profiles and expected temporal lag. We also account for readout complexity separately from encoder cost.
- **An empirical assessment across encoder families.** We report comparisons with ten baselines on five benchmarks and examine matched-backbone and component ablations. The results identify encoder-dependent gains and clarify where a more flexible temporal readout does, and does not, improve the reported ranking performance.

2 Related Work

Sequence encoders. Self-attention models such as SASRec (Kang & McAuley, 2018), BERT4Rec (Sun et al., 2019), and FEARec (Du et al., 2023) learn interactions among sequence positions. Mamba (Gu & Dao, 2023) uses a multidimensional state and input-dependent selection, supporting several implicit memory scales. Mamba4Rec (Liu et al., 2024a) and SIGMA (Liu et al., 2025) adapt this approach to sequential recommendation. Our motivation is not that these encoders lack multi-scale capacity, but that an explicit timestamp-conditioned readout can be controlled separately from their internal dynamics.

Time-aware sequential recommenders. TiSASRec (Li et al., 2020) introduces relative-interval embeddings into self-attention. MEANTIME (Cho et al., 2020) uses several temporal embeddings and attention mechanisms. Contextualized Temporal Attention (CTA) combines parameterised temporal kernels with context-dependent selection, making it a close antecedent to multi-kernel temporal weighting. Gradformer (Liu et al., 2024b) studies exponential decay masks in graph Transformers. MARS builds on these principles through a particular post-encoder factorisation: normalised recency heads with user-conditioned effective rates, followed by adaptive fusion. The distinction is the placement and parameterisation of the readout, rather than the first use of multiple temporal kernels or contextual routing.

Aggregation strategies. Last-position readout is common in SASRec, FEARec and Mamba4Rec; BERT4Rec obtains its prediction representation from a masked position. Time-aware encoders can propagate temporal information into these representations, while gated architectures provide additional selection mechanisms. MARS instead takes hidden states and observed timestamps as separate inputs to its readout, returning a vector compatible with the existing item scorer. Table 1 summarises where temporal modelling enters representative methods. Comparisons with last-position readouts isolate the aggregation question more directly than comparisons between complete architectures.

Table 1: Representative temporal modelling choices. "Placement" describes where temporal weighting or memory dynamics enter the architecture; it is not a count of the model's implicit time-scales.

Method	Temporal mechanism	Placement	Encoder
SASRec	positional attention	encoder	Transformer
BERT4Rec	positional attention	encoder	Transformer
FEARec	frequency-enhanced mixing	encoder	Transformer
TiSASRec	relative-interval embeddings	encoder	Transformer
MEANTIME	multiple temporal embeddings	encoder	Transformer
CTA	context-selected temporal kernels	attention	self-attention
Mamba4Rec	selective state dynamics	encoder	Mamba
SIGMA	gated state-space dynamics	encoder	Mamba
MARS	user-conditioned recency heads	post-encoder	Trans./Mamba

3 METHOD: MARS

3.1 PROBLEM DEFINITION

Let $\mathcal{U}$ and $\mathcal{V}$ denote the user and item sets. A user u has an observed interaction history $S_u = \{(v_i, t_i)\}_{i=1}^{L_u}$, with item v_i and timestamp t_i. We predict the next item using only this history. The retained suffix contains $\ell_u = \min(L_u, L)$ events and is padded to length L. Write $\mathcal{I}_u$ for its valid positions, j_u for the last valid position, and t_{last} for the latest observed timestamp. For valid positions, $\Delta t_i = t_{\text{last}} - t_i \geq 0$; no target timestamp is required. The encoder returns $\mathbf{H} \in \mathbb{R}^{L \times d}$, and $m_i = 1$ on $\mathcal{I}_u$ and 0 on padding. We assume a nonempty retained history.

3.2 OVERVIEW

The architecture of MARS is shown in Figure 1. A sequence encoder transforms the item history into hidden states $\mathbf{H} \in \mathbb{R}^{L \times d}$, which the post-encoder MARS module processes in three stages: per-user multi-rate recency aggregation that produces K time-scale summaries (Section 3.3), context-adaptive fusion that combines them into a single user vector (Section 3.4), and joint optimisation of all parameters with two auxiliary regularisers (Section 3.5). Two encoder instantiations are studied under an identical protocol: MARS-T stacks N Transformer blocks with causal masking on top of a learnable positional embedding, inheriting the configuration of SASRec (Kang & McAuley, 2018); MARS-M replaces the Transformer blocks with N Mamba blocks following Mamba4Rec (Liu et al., 2024a), with a feed-forward sub-layer per block. We denote $\mathbf{h}_{\text{last}} \in \mathbb{R}^d$ the row of $\mathbf{H}$ at the user's last unmasked position j_u.

3.3 MULTI-RATE RECENCY AGGREGATION

Per-head decay weighting. The elapsed times $\{\Delta t_i\}_{i=1}^{L}$ are first log-compressed to

$$\tilde{\Delta} t_i \;=\; \log\bigl(1 + \Delta t_i / \tau_0\bigr), \tag{1}$$

with τ_0 fixed at one hour. The compression reduces the dynamic range of inter-event gaps spanning several orders of magnitude; padding positions inherit $\tilde{\Delta} t_i = 0$. Each of the K decay heads carries two scalar parameters, a base rate $\lambda_k = \mathrm{softplus}(\tilde{\lambda}_k)$ and a temperature $\tau_k = \mathrm{softplus}(\tilde{\tau}_k) + 0.1$. The unconstrained parameter $\tilde{\lambda}_k$ is initialised on a linspace from -1 to $+2$, so the heads begin from a diverse decomposition of the time axis.

User-conditioned modulation. To inject per-user heterogeneity without per-user parameters, the base rate of head k is modulated by a bounded projection of the user summary $\mathbf{h}_{\text{last}}$,

$$\lambda_k(u) \;=\; \lambda_k \cdot \exp\bigl(\sigma \cdot \tanh(\mathbf{W}_\lambda \, \mathbf{h}_{\text{last}})_k\bigr), \tag{2}$$

with $\mathbf{W}_\lambda \in \mathbb{R}^{K \times d}$ and $\sigma = 0.5$ bounding the modulation magnitude. Zero-initialising $\mathbf{W}_\lambda$ gives the unmodulated initial regime $\lambda_k(u) = \lambda_k$.

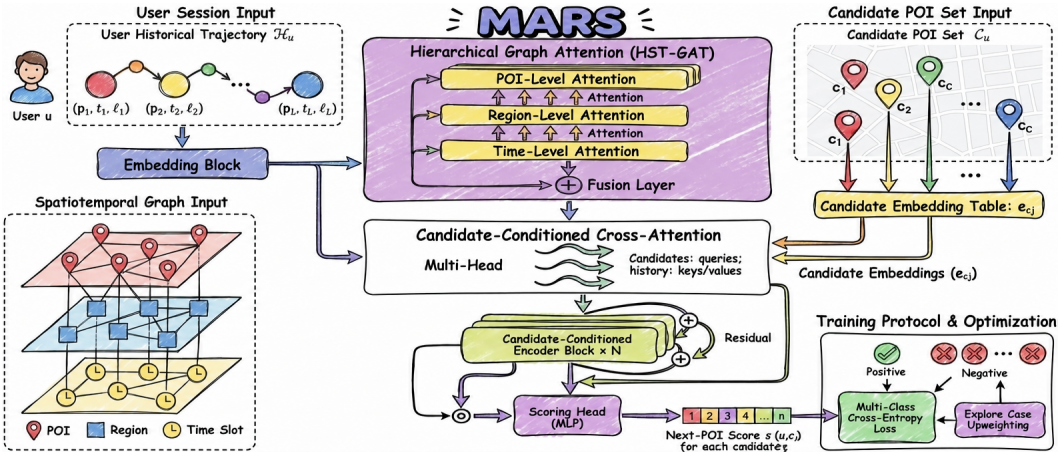


Figure 1: MARS architecture. A sequence encoder produces hidden states $\mathbf{H}$. MARS aggregates them into K recency-conditioned summaries via learnable rates $\{\lambda_k\}_{k=1}^{K}$ that are modulated per user, then fuses them into the user vector $\mathbf{h}_u$ through a context-adaptive gate.

Aggregation. The K summaries are obtained by softmax-normalising the per-head decay logits over the unmasked positions and aggregating $\mathbf{H}$ accordingly,

$$w_{k,i}^{(u)} = \frac{\exp(-\lambda_k(u)\,\tilde{\Delta}t_i/\tau_k)\,m_i}{\sum_{j=1}^{L}\exp(-\lambda_k(u)\,\tilde{\Delta}t_j/\tau_k)\,m_j}, \qquad \mathbf{Z}_k^{(u)} = \sum_{i=1}^{L} w_{k,i}^{(u)}\,\mathbf{h}_i. \tag{3}$$

The matrix-vector formulation $\mathbf{Z} = \mathbf{w}\,\mathbf{H} \in \mathbb{R}^{K\times d}$ collapses the entire aggregation into a single batched matrix multiplication. The choice of a K-rate exponential mixture as the aggregation kernel is motivated by the following approximation result.

Proposition 1 (Approximation of Laplace-mixture kernels)**.** *Let $0 \le S < \infty$ and let ν be a probability measure on $[0,\infty)$. Define the kernel $\phi(s) = \int e^{-rs}\,\nu(dr)$ for $s \in [0,S]$. For every $\varepsilon > 0$, there exist finitely many positive rates $r_1,\dots,r_K$ and non-negative weights $\beta_1,\dots,\beta_K$ with $\sum_k \beta_k = 1$ such that*

$$\sup_{s\in[0,S]}\left|\phi(s) - \sum_{k=1}^{K}\beta_k e^{-r_k s}\right| \le \varepsilon. \tag{4}$$

Appendix C proves the claim by truncating and discretising ν. This result concerns a restricted kernel family in the compressed coordinate $s = \log(1+\Delta t/\tau_0)$, rather than all monotone functions of elapsed time. In Eq. equation 3, the effective rate is $r_k(u) = \lambda_k(u)/\tau_k$, so each unnormalised head is exponential in s and has the form $(1+\Delta t/\tau_0)^{-r_k(u)}$ in physical time. For a fixed history, let $Z_k^{\mathrm{norm}} = \sum_{i:m_i=1} e^{-r_k(u)s_i}$. For a valid position i, a gate with weights α_k gives effective position weights $q_i = \sum_k \alpha_k e^{-r_k(u)s_i}/Z_k^{\mathrm{norm}}$. Thus the unnormalised mixture coefficients are proportional to $\alpha_k/Z_k^{\mathrm{norm}}$, not simply α_k. The proposition motivates the kernel bank; it does not guarantee that the learned gate attains every approximating mixture or that training recovers a data-generating kernel.

3.4 Context-Adaptive Fusion

The K summaries $\{\mathbf{Z}_k^{(u)}\}_{k=1}^{K}$ are fused through a gate conditioned on a context vector $\mathbf{c} = [\mathrm{tEnc}(0);\ \mathbf{e}_{v_{j_u}};\ \mathbf{h}_{\text{last}}]$, where tEnc is a two-layer MLP applied to a constant zero input and $\mathbf{e}_{v_{j_u}} \in \mathbb{R}^d$ is the embedding of the most recent item:

$$g_k = \mathbf{v}^\top \tanh\big(\mathbf{W}_g[\mathbf{Z}_k^{(u)};\,\mathbf{c}]\big), \quad \boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{g}/\tau_\alpha), \quad \mathbf{h}_u = \mathbf{h}_{\text{last}} + \sum_{k=1}^{K}\alpha_k\,\mathbf{Z}_k^{(u)}, \tag{5}$$

with learnable $\mathbf{W}_g$ and $\mathbf{v}$ and learnable temperature $\tau_\alpha > 0$, whose reported default value is 1. Because the latest observed event defines the time origin, its elapsed time is zero: $\mathrm{tEnc}(0)$ is a shared

learned offset, not a varying time feature. Temporal variation enters through the head summaries $\mathbf{Z}_k^{(u)}$. The residual preserves a direct last-position path, but does not guarantee improved ranking or allow the softmax weights to vanish together. Algorithm A1 gives the forward pass.

The aggregation is linear in retained sequence length for fixed hidden size and head count. The full readout also includes small dense projections.

Proposition 2 (Readout complexity)**.** *Let the gate have hidden width h_g and context dimension $\mathcal{O}(d)$. Given encoder states and the shared offset* $\mathrm{tEnc}(0)$*, MARS requires $\mathcal{O}(LdK + Kdh_g + dK)$ forward time and $\mathcal{O}(LK + dK + Kh_g)$ intermediate storage, excluding model parameters and the supplied states. The shared offset can be computed once for inference and contributes an additional cost independent of L when recomputed during training. For fixed d, h_g and K, the readout is linear in L.*

Proof. Computing decay logits and normalisers costs $\mathcal{O}(LK)$, and $\mathbf{wH}$ costs $\mathcal{O}(LdK)$. Modulation costs $\mathcal{O}(dK)$, the dense gate costs $\mathcal{O}(Kdh_g)$, and fusion costs $\mathcal{O}(dK)$. Storing weights, summaries and gate activations gives the stated bound. □

This is incremental readout cost. MARS-T retains the Transformer's quadratic attention term, while MARS-M retains the sequence scaling of its Mamba encoder. Direct pairwise evaluation of the training-time JSD regulariser adds $\mathcal{O}(LK^2)$ work; it is not needed at inference.

3.5 Optimisation

The standard sequential cross-entropy loss is augmented with two auxiliary regularisers,

$$\mathcal{L} \;=\; \mathcal{L}_{\text{CE}} \;+\; \eta_{\text{div}}\,\mathcal{L}_{\text{div}} \;+\; \eta_{\text{bal}}\,\mathcal{L}_{\text{bal}}, \tag{6}$$

with default coefficients $\eta_{\text{div}} = \eta_{\text{bal}} = 10^{-2}$. The diversity term $\mathcal{L}_{\text{div}} = -\sum_{k\neq k'} \mathrm{JSD}(w_{k,\cdot}^{(u)} \| w_{k',\cdot}^{(u)})$ is the negative pairwise Jensen–Shannon divergence of the per-head attention distributions, and the load-balance term $\mathcal{L}_{\text{bal}} = K\sum_{k=1}^{K} \bar{\alpha}_k^2$, with $\bar{\alpha}_k$ the batch mean of the k-th gate weight, is a batch-level balance penalty that discourages aggregate routing concentration. Neither penalty guarantees distinct heads or uniform per-user routing.

The diversity term encourages different attention profiles across heads. The following property characterises this effect for a fixed history and also identifies its limits.

Proposition 3 (Separation of effective-rate attention heads)**.** *Fix a non-empty history with valid compressed lags $s_1, \ldots, s_n$, and define $p_i(r) = e^{-rs_i} / \sum_{j=1}^{n} e^{-rs_j}$ for finite $r > 0$. If at least two lags differ, the map $r \mapsto \mathbf{p}(r)$ is injective, and*

$$\frac{d}{dr}\sum_i p_i(r)s_i = -\operatorname{Var}_{\mathbf{p}(r)}(s) < 0. \tag{7}$$

Consequently, for any $K \geq 2$ finite positive effective rates, $\sum_{k\neq k'} \mathrm{JSD}(\mathbf{p}(r_k)\|\mathbf{p}(r_{k'})) = 0$ if and only if all r_k coincide. If all valid lags are equal, every head is uniform regardless of its rate.

Sketch. For $s_i \neq s_j$, the log probability ratio is $\log[p_i(r)/p_j(r)] = -r(s_i - s_j)$, which determines r uniquely. Differentiating the expected lag gives the negative variance. Non-negativity of JSD, with equality exactly for identical distributions, then gives the final statement. Appendix D provides details. □

The proposition shows that distinct effective rates yield distinct attention profiles whenever a history contains temporal variation, and larger rates put more mass on recent interactions in the sense of a smaller expected lag. It does not establish recovery of physical decay rates or exclude partial head duplication: a positive total divergence only requires at least one pair of different profiles. Moreover, λ_k and τ_k are not separately identifiable because a common rescaling leaves their ratio unchanged. The diversity penalty is therefore an optimisation preference, not a guarantee that every trained head remains distinct.

3.6 BACKBONE SELECTION

Given a dataset, we summarise the observed encoder preference with a fixed rule based on mean history length $\bar{L}$:

$$\text{backbone}(\mathcal{D}) = \begin{cases} MARS-T & \bar{L}(\mathcal{D}) < 50, \\ MARS-M & \bar{L}(\mathcal{D}) \geq 50. \end{cases} \tag{8}$$

For the present benchmark comparison, the dataset statistics in Table 2 place all four sparse datasets below the threshold and ML-1M above it. Both instantiations are reported separately to make the encoder comparison explicit.

The rule is a descriptive dataset-level heuristic motivated by this benchmark set, not a learned or independently validated routing policy. The observed length gap leaves the threshold weakly determined: many values induce the same assignments. For a new deployment, the corresponding statistic must be computed from available training histories; the current experiments do not establish how the rule transfers to intermediate regimes. Attention-entropy summaries in Appendix H provide additional descriptive context, without identifying a causal switching boundary.

4 EXPERIMENTS

4.1 SETTINGS

Datasets. We evaluate on five widely-used benchmarks summarised in Table 2, namely the three Amazon categories Beauty, Sports, and Games released by He & McAuley (2016), MovieLens-1M (Harper & Konstan, 2015), and Yelp (Yelp Inc.). Beauty, Sports, Games and Yelp belong to the sparse, short-history regime, with average sequence length below 13; ML-1M is a dense regime with sequences nearly twenty times longer. These datasets provide contrasting history-length regimes, with only one dense benchmark; they do not cover the full range of deployment conditions.

Table 2: Statistics of the benchmarks.

Dataset	#Users	#Items	#Inters.	Sparsity	Avg.
Beauty	22,363	12,101	198,502	99.93%	8.88
Sports	35,598	18,357	296,337	99.95%	8.32
Games	94,762	25,612	814,586	99.97%	8.60
ML-1M	6,040	3,416	999,611	95.16%	165.5
Yelp	233,247	109,062	2,918,699	99.99%	12.51

Baselines. Ten representative baselines spanning RNN, attention-based, time-aware, session-based, and state-space families are re-run under our unified RecBole (Zhao et al., 2021) framework. They are the recurrent GRU4Rec (Hidasi et al., 2016), the attention-augmented recurrent NARM (Li et al., 2017), the bidirectional Transformer BERT4Rec (Sun et al., 2019), the causal Transformer SASRec (Kang & McAuley, 2018), the time-aware Transformer TiSASRec (Li et al., 2020), the frequency-domain FEARec (Du et al., 2023), the consistent-representation session model CORE (Hou et al., 2022), and three state-space models, namely Mamba4Rec (Liu et al., 2024a), EchoMamba4Rec (denoted Echo in tables) (Wang et al., 2024), and SIGMA (Liu et al., 2025). The reported comparison uses re-trained baselines rather than values quoted from their publications. Shared splits reduce one source of variation; matched-backbone ablations are still needed to separate readout effects from architectural differences.

Implementation details. All Transformer-based models share the SASRec configuration of d=64, two attention blocks, inner dimension 256, and dropout 0.5 on Beauty, Sports, Yelp and 0.2 on Games, ML-1M. Mamba-based models use the Mamba4Rec defaults of d_{state}=32, d_{conv}=4, expand=2. Optimisation uses Adam (Kingma & Ba, 2015) at learning rate 10^{-3} with batch size 2048 for up to 200 epochs and early stopping on validation NDCG@10 with patience 20. MARS uses K=4 decay heads on the four sparse benchmarks and K=8 on ML-1M to match the longer sequences. All experiments are performed on a single NVIDIA V100 GPU. The main comparison reports mean±std over five seeds for each MARS configuration; the baselines use one seed each. Seed-level correspondence for the separate ablation and diagnostic summaries is not established by the available records.

Table 3: Comparison results ($\times 100$). MARS reports mean$\pm$std over five seeds. Backbone is selected by $\bar{L}$, with † marking MARS-T ($\bar{L} < 50$) and ‡ marking MARS-M ($\bar{L} \geq 50$). **Best**, second.

Dataset	Metric	GRU4Rec	BERT4Rec	NARM	SASRec	TiSASRec	FEARec	CORE	Mamba4Rec	Echo	SIGMA	MARS-T	MARS-M	MARS
Beauty	HR@10	5.82	3.94	6.03	8.04	8.78	7.83	6.49	6.62	6.93	6.03	$8.84_{\pm 0.14}$	$8.02_{\pm 0.23}$	**8.84**†
	NDCG	3.25	1.95	3.45	4.10	**4.45**	4.08	2.59	4.12	4.30	3.43	$4.40_{\pm 0.06}$	$4.13_{\pm 0.08}$	4.40†
	MRR	2.47	1.36	2.67	2.90	3.12	2.92	1.44	3.36	**3.50**	2.64	$3.04_{\pm 0.06}$	$2.94_{\pm 0.04}$	3.04†
Sports	HR@10	3.06	1.66	3.12	4.53	4.84	4.60	3.24	3.50	3.40	3.20	$5.22_{\pm 0.08}$	$4.73_{\pm 0.19}$	**5.22**†
	NDCG	1.58	0.81	1.60	2.14	2.30	2.19	1.26	2.08	2.08	1.77	$2.43_{\pm 0.04}$	$2.26_{\pm 0.07}$	**2.43**†
	MRR	1.13	0.56	1.14	1.41	1.52	1.44	0.69	1.65	**1.67**	1.33	$1.57_{\pm 0.03}$	$1.51_{\pm 0.05}$	1.57†
Games	HR@10	7.80	6.69	10.15	8.86	**12.07**	8.64	8.84	8.02	8.02	8.10	$12.07_{\pm 0.16}$	$12.01_{\pm 0.42}$	**12.07**†
	NDCG	4.05	3.49	5.24	4.34	**5.90**	4.20	3.61	4.21	4.18	4.20	$5.51_{\pm 0.09}$	$5.62_{\pm 0.11}$	5.51†
	MRR	2.92	2.53	3.77	2.97	**4.03**	2.84	2.06	3.05	3.01	3.03	$3.52_{\pm 0.07}$	$3.68_{\pm 0.07}$	3.52†
ML-1M	HR@10	29.37	23.11	29.37	29.62	24.90	28.05	15.55	30.28	31.04	31.79	$29.04_{\pm 0.70}$	$32.80_{\pm 0.45}$	**32.80**‡
	NDCG	16.90	11.87	16.58	16.61	13.42	15.63	6.79	17.72	18.00	18.63	$16.39_{\pm 0.50}$	$18.80_{\pm 0.32}$	**18.80**‡
	MRR	13.10	8.48	12.67	12.65	9.93	11.86	4.18	13.88	14.03	**14.60**	$12.52_{\pm 0.45}$	$14.52_{\pm 0.29}$	14.52‡
Yelp	HR@10	2.98	2.38	3.42	5.41	5.62	4.79	6.14	4.35	3.79	3.80	$6.45_{\pm 0.04}$	$5.97_{\pm 0.04}$	**6.45**†
	NDCG	1.65	1.21	1.89	3.64	3.84	3.42	3.65	2.55	2.13	2.12	$4.22_{\pm 0.01}$	$3.89_{\pm 0.01}$	**4.22**†
	MRR	1.25	0.86	1.42	3.10	3.30	3.00	2.89	2.00	1.63	1.61	$3.53_{\pm 0.01}$	$3.25_{\pm 0.02}$	**3.53**†

4.2 Comparison

Table 3 reports both instantiations and ten baselines. The selected model has the highest reported HR@10 on each dataset, including a displayed tie with TiSASRec on Games. The following comparisons concern the table's point estimates, not significance-tested superiority.

(i) Strong sparse-data retrieval. Relative to the strongest content-only Transformer on each sparse benchmark, MARS-T improves HR@10 by 9.95%, 13.48%, 36.23% and 19.22%, averaging 19.72%. Against the time-aware TiSASRec, the corresponding average is 5.83% from the displayed values. The choice of reference matters: on Games, MARS-T and TiSASRec tie at 12.07 HR@10, while TiSASRec has higher NDCG and MRR. MARS-T leads NDCG on Sports and Yelp, whereas Echo leads MRR on Beauty and Sports. Thus the strongest empirical case is improved top-10 retrieval, with metric-dependent trade-offs.

(ii) A different encoder preference on ML-1M. MARS-M reaches 32.80 HR@10 and 18.80 NDCG@10, compared with SIGMA's 31.79 and 18.63, giving relative differences of 3.18% and 0.91%. Its MRR is slightly lower (14.52 versus 14.60). The reported forward-cost estimate is 41.85% below SIGMA's. These are complete-model comparisons; the matched-backbone ablation is the more direct estimate of the readout's contribution. On sparse datasets, MARS-T has the higher reported HR@10, although the Games difference is small relative to the reported variability.

(iii) Scope of the repeated runs. The five-seed standard deviations quantify variation within each MARS configuration. Baselines have one reported seed, so these deviations do not establish the variance of between-method differences. In particular, the Beauty HR@10 margin over TiSASRec and the ML-1M NDCG margin over SIGMA should be read as modest point-estimate differences, not confirmed statistical gains.

4.3 Ablation Study

Table 4: Reported MARS-T ablation summary (HR@10, $\times 100$). Full-model reference values differ from the expanded summary in Appendix E; component effects below are calculated within this table. Seed pairing is not established by the available summaries.

Variant	Beauty	Sports	Games	ML-1M	Yelp
Full MARS-T	8.91	5.20	12.24	29.30	6.40
Backbone only (w/o MARS)	8.53	4.89	11.75	30.33	5.36
b1 (w/o real Δt)	8.69	5.16	12.32	29.82	6.49
b2 (w/o $\lambda(u)$)	8.94	5.21	12.32	29.25	6.39
b3 (w/o diversity)	8.90	5.21	12.14	29.22	6.37
b4 (w/o balance)	8.88	5.12	12.26	28.41	6.33
Single head (K=1)	8.80	5.27	12.23	27.45	6.42

(i) Matched-backbone gains on sparse data. Within Table 4, adding the full readout changes HR@10 by $+4.45\%$ on Beauty, $+6.34\%$ on Sports, $+4.17\%$ on Games and $+19.40\%$ on Yelp. These differences are smaller than some complete-model comparisons in Table 3, underscoring the distinction between readout gains and gains relative to a separately trained baseline. They support the readout as a useful addition in these reported configurations, without isolating a single responsible component. **(ii) Encoder**

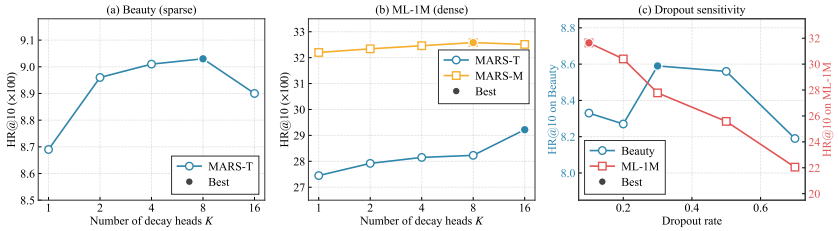


Figure 2: Hyperparameter sensitivity of MARS. (a) K on Beauty, (b) K on ML-1M for MARS-T and MARS-M, and (c) dropout on Beauty and ML-1M.

dependence on dense histories. On ML-1M, the Transformer backbone-only value is 30.33, above the full readout's 29.30. In the Mamba ablation (Table A2), the full value is 32.80 versus 32.40 for the matched backbone, a 0.40-point increment. This contrast motivates considering the encoder and readout jointly. It does not imply that dense histories universally favour Mamba or that attention concentration causes the difference. **(iii) Component effects are heterogeneous.** Removing real elapsed time lowers Beauty HR@10 from 8.91 to 8.69, but raises the reported Games and Yelp values. A single head is competitive on the sparse datasets and exceeds Full on Sports and Yelp. User conditioning and diversity regularisation likewise do not improve every cell. These observations favour a modest interpretation: the combined readout can help, while the benefit of additional temporal flexibility depends on the setting. The auxiliary penalties express useful modelling preferences, not empirically established necessities across all benchmarks.

4.4 Hyperparameter Sensitivity

(i) Small head banks can suffice on sparse data. In the Beauty sweep, $K = 1$ reaches 8.69 HR@10 and $K = 8$ reaches 9.03; the single-head result is 96.2% of the best value. A larger bank therefore offers a modest improvement in this sweep. The approximation result describes a representational family, not a prediction that increasing K must improve ranking. **(ii) Head-count sensitivity depends on the encoder.** On ML-1M, MARS-T increases from 27.45 at $K = 1$ to 29.22 at $K = 16$, while MARS-M ranges from 32.20 to 32.58, peaking at $K = 8$. The smaller Mamba variation is compatible with greater overlap between its encoded information and the readout, but these curves alone do not identify that mechanism. The reported defaults remain $K = 4$ for sparse datasets and $K = 8$ for ML-1M. **(iii) Regularisation materially affects performance.** The dropout sweep peaks at 0.3 on Beauty and 0.1 on ML-1M. On ML-1M, changing dropout from 0.1 to 0.5 reduces HR@10 from 31.64 to 25.58. These are sensitivity results, not evidence of uniform superiority across the grid. The sweep and main-comparison summaries are reported separately; their values should not be treated as matched-seed estimates without configuration-level correspondence.

4.5 Efficiency Analysis

Table 5 records the reported parameter, forward-operation, timing and memory summaries. Relative to SASRec, MARS-T has 2.1% more parameters on Beauty and 5.4% more on ML-1M. Relative to SIGMA, MARS-M has 41.4% fewer MFLOPs on Beauty and 41.8% fewer on ML-1M. These latter comparisons are between complete architectures: MARS-M is deeper than the standard Mamba4Rec baseline, so its extra parameters cannot all be assigned to the readout. The reported MARS-T MFLOPs are lower than SASRec's, although the described post-encoder operation does not remove attention. This difference cannot be explained by the aggregation formula and requires a common implementation-level counting scope before it is interpreted as an architectural saving. Similarly, the recorded timing and memory values depend on the profiling setup and do not establish asymptotic or deployment-wide dominance. The supported computational distinction is that the readout adds a cost linear in sequence length for fixed d and K; it leaves the encoder's scaling intact.

4.6 Subgroup Analysis: Where Does MARS Help Most?

Table 6 retains the reported length-stratified summaries. They are exploratory diagnostics rather than a decomposition of Table 3: the reported subgroup and aggregate values do not reconcile

Table 5: Efficiency on Beauty and ML-1M (NVIDIA V100).

Model	Beauty				ML-1M			
	#P (K)	MFLOPs	Time (s)	Mem (G)	#P (K)	MFLOPs	Time (s)	Mem (G)
GRU4Rec	857	4.14	0.044	1.80	301	16.56	0.035	7.07
BERT4Rec	894	5.32	0.907	1.19	339	22.70	0.503	11.97
SASRec	878	5.10	0.197	1.22	332	21.83	0.313	12.56
FEARec	878	12.43	0.726	1.40	332	49.72	0.726	5.93
Mamba4Rec	847	1.68	0.124	0.89	291	6.71	0.097	3.09
Echo	864	2.52	0.348	1.38	308	10.09	0.276	5.05
SIGMA	905	5.80	0.300	1.31	349	23.18	0.251	5.12
MARS-T	896	5.05	0.151	1.05	350	20.27	0.173	4.56
MARS-M	937	3.40	0.237	0.89	382	13.48	0.181	3.11

as one weighted evaluation. For example, all Sports SASRec buckets are at most 4.13, below the aggregate value 4.53, and all Games MARS buckets are at most 11.31, below 12.07. Without matched prediction records, these summaries cannot support an aggregate-gain or causal-mechanism claim.

(i) Sparse benchmarks. The reported Yelp differences increase across length buckets, while Games includes a negative short-history difference. These contrasting patterns suggest a useful diagnostic question about heterogeneity, but do not establish a common length-dependent benefit. **(ii) Dense ML-1M.** The two encoder variants display different patterns in the reported buckets. Their interpretation additionally depends on whether bucketing uses raw or truncated history length, as discussed in Appendix G. **(iii) Scope.** We therefore base the main performance discussion on Table 3 and the within-table ablation contrasts, rather than infer a temporal mechanism from the subgroup summaries.

Table 6: Subgroup HR@10 ($\times 100$) by user history length L_u. Each cell: baseline $\rightarrow$ MARS ($\Delta\%$). The baseline is SASRec for the four sparse benchmarks and the ML-1M, MARS-T row, and Mamba4Rec for the ML-1M, MARS-M row. Red marks $\Delta < 0$. Bucket definitions are discussed in Appendix G; complete bucket counts and matched-checkpoint correspondence are unavailable in the available summaries.

Dataset	Short	Medium	Long	Very long
Beauty	6.57 $\rightarrow$ 7.98 (+21.4)	7.66 $\rightarrow$ 8.40 (+9.7)	12.89 $\rightarrow$ 14.49 (+12.4)	20.26 $\rightarrow$ 18.95 (-6.5)
Sports	4.13 $\rightarrow$ 5.30 (+28.5)	3.91 $\rightarrow$ 4.97 (+27.2)	3.25 $\rightarrow$ 4.35 (+33.7)	2.25 $\rightarrow$ 5.62 (+149.8)
Games	12.05 $\rightarrow$ 11.31 (-6.1)	10.91 $\rightarrow$ 10.94 (+0.3)	10.09 $\rightarrow$ 10.25 (+1.6)	8.10 $\rightarrow$ 8.91 (+10.0)
Yelp	5.69 $\rightarrow$ 6.93 (+21.8)	5.20 $\rightarrow$ 6.50 (+25.0)	4.36 $\rightarrow$ 5.49 (+25.9)	3.64 $\rightarrow$ 4.72 (+29.7)
ML-1M (MARS-T)	7.92 $\rightarrow$ 8.04 (+1.5)	18.43 $\rightarrow$ 19.13 (+3.8)	27.51 $\rightarrow$ 27.85 (+1.2)	30.85 $\rightarrow$ 29.96 (-2.9)
ML-1M (MARS-M)	8.21 $\rightarrow$ 8.68 (+5.7)	18.86 $\rightarrow$ 20.32 (+7.7)	28.42 $\rightarrow$ 30.97 (+9.0)	31.45 $\rightarrow$ 34.01 (+8.1)

4.7 Discussion

MARS exposes temporal readout as a modular modelling choice, with effective decay rates controlling recency weighting and adaptive fusion combining the resulting summaries. The reported gains depend on the encoder and dataset, while the backbone-selection rule remains empirical. Appendix J discusses the operator's interpretation, evidence limitations, and broader impact in detail.

5 Conclusion

We presented MARS as an explicit temporal readout for sequential recommendation. User-conditioned effective rates select recency profiles over encoder states, and adaptive fusion combines them with a direct last-position path. A restricted kernel-approximation result and an effective-rate characterisation explain the operator's scope without asserting recovery of latent behavioural parameters. The reported comparisons show competitive top-10 retrieval and encoder-dependent gains, with heterogeneous effects in component ablations. The broader lesson is to treat temporal readout as a separately controlled modelling choice whose value should be assessed against its matched encoder.

AI USE STATEMENT

AI tools assisted with language checking and polishing. No additional model-training or inference experiments were performed as part of these revisions. The authors take responsibility for the accuracy of the reported results, citations, and final manuscript.

ETHICS STATEMENT

This work studies time-aware sequential recommendation using existing benchmark datasets. The reported results concern offline ranking performance under the evaluated settings and do not establish fairness, privacy protection, or effectiveness in deployment. User interaction histories and timestamps can reveal sensitive behavioral patterns, and practical applications require appropriate consent and data protection measures. Recommendation systems may also reinforce popularity bias and narrow users' exposure to content; these effects are not assessed in this study.

REPRODUCIBILITY STATEMENT

Code is available at https://anonymous.4open.science/r/MARSS. The repository includes a reference training configuration, saved prediction records, and scripts for recomputing statistical summaries. Experimental settings and evaluation protocols are documented in the main text and appendix.

REFERENCES

Sung Min Cho, Eunhyeok Park, and Sungjoo Yoo. MEANTIME: Mixture of attention mechanisms with multi-temporal embeddings for sequential recommendation. In *Fourteenth ACM Conference on Recommender Systems*, pp. 515–520, 2020. doi: 10.1145/3383313.3412216.

Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S. Sheng. Frequency enhanced hybrid attention network for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 78–88, 2023. doi: 10.1145/3539618.3591689.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. URL https://arxiv.org/abs/2312.00752.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pp. 507–517, 2016. doi: 10.1145/2872427.2883037.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. URL https://openreview.net/forum?id=yoffK5KZSgQ.

Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. CORE: Simple and effective session-based recommendation within consistent representation space. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 1796–1801, 2022. doi: 10.1145/3477495.3531955.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018. doi: 10.1109/ICDM.2018.00035.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 322–330, 2020. doi: 10.1145/3336191.3371786.

Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*, pp. 1419–1428, 2017. doi: 10.1145/3132847.3132926.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. Mamba4Rec: Towards efficient sequential recommendation with selective state space models. In *Second International Workshop on Resource-Efficient Learning for Knowledge Discovery (RelKD), held in conjunction with KDD 2024*, 2024a. URL https://arxiv.org/abs/2403.03900.

Chuang Liu, Zelin Yao, Yibing Zhan, Xueqi Ma, Shirui Pan, and Wenbin Hu. Gradformer: Graph transformer with exponential decay. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2171–2179, 2024b. doi: 10.24963/ijcai.2024/240.

Ziwei Liu, Qidong Liu, Yejing Wang, Wanyu Wang, Pengyue Jia, Maolin Wang, Zitao Liu, Yi Chang, and Xiangyu Zhao. SIGMA: Selective gated mamba for sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12264–12272, 2025. doi: 10.1609/aaai.v39i12.33336. URL https://ojs.aaai.org/index.php/AAAI/article/view/33336.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019. doi: 10.1145/3357384.3357895.

Yuda Wang, Xuxin He, and Shengxin Zhu. EchoMamba4Rec: Harmonizing bidirectional state space models with spectral filtering for advanced sequential recommendation. *arXiv preprint arXiv:2406.02638*, 2024. URL https://arxiv.org/abs/2406.02638.

Yelp Inc. Yelp open dataset. https://www.yelp.com/dataset.

Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. RecBole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM)*, pp. 4653–4664, 2021. doi: 10.1145/3459637.3482016.

A Appendix

B MARS Forward Pass

The full MARS forward pass for a single user is given in Algorithm A1. Batching across users is performed via a single matrix multiplication on Step 5, with all other operations being pointwise or small per-user MLPs.

Algorithm A1 MARS forward pass.

Require: Encoder outputs $\mathbf{H} \in \mathbb{R}^{L\times d}$, $\mathbf{h}_{\text{last}} \in \mathbb{R}^d$, mask $\mathbf{m} \in \{0,1\}^L$, elapsed times $\Delta t \in \mathbb{R}^L_{\geq 0}$, last item index v_{j_u}, embedding table $\mathbf{E}$; learnable rates $\boldsymbol{\lambda}, \boldsymbol{\tau} \in \mathbb{R}^K_{>0}$, modulation $\mathbf{W}_\lambda$, fusion $(\mathbf{W}_g, \mathbf{v}, \tau_\alpha)$, time encoder tEnc; hyperparameters σ, τ_0.

Ensure: user representation $\mathbf{h}_u \in \mathbb{R}^d$.

1: $\tilde{\Delta}t_i \leftarrow \log\big(1 + \Delta t_i/\tau_0\big) \quad \forall i$ ▷ compress gaps

2: $\boldsymbol{\delta} \leftarrow \tanh(\mathbf{W}_\lambda\, \mathbf{h}_{\text{last}}), \quad \boldsymbol{\lambda}^{(u)} \leftarrow \boldsymbol{\lambda} \odot \exp(\sigma\, \boldsymbol{\delta})$ ▷ user-conditioned rates

3: $\ell_{k,i} \leftarrow -\lambda_k^{(u)}\, \tilde{\Delta}t_i/\tau_k \quad \forall k, i$

4: Set $\ell_{k,i} \leftarrow -\infty$ wherever $m_i = 0$; $\quad w_{k,i} \leftarrow \mathrm{softmax}_i(\ell_{k,i})$

5: $\mathbf{Z} \leftarrow \mathbf{w}\,\mathbf{H} \in \mathbb{R}^{K\times d}$ ▷ $\mathcal{O}(LdK)$

6: $\mathbf{c} \leftarrow [\,\mathrm{tEnc}(0);\ \mathbf{e}_{v_{j_u}};\ \mathbf{h}_{\text{last}}\,]$

7: $g_k \leftarrow \mathbf{v}^\top \tanh\big(\mathbf{W}_g[\mathbf{Z}_k; \mathbf{c}]\big), \quad \boldsymbol{\alpha} \leftarrow \mathrm{softmax}(\mathbf{g}/\tau_\alpha)$

8: $\mathbf{h}_u \leftarrow \mathbf{h}_{\text{last}} + \sum_{k=1}^K \alpha_k\, \mathbf{Z}_k$ ▷ fuse + residual

9: **return** $\mathbf{h}_u$

C Proof of Proposition 1

Proof. If $S = 0$, every mixture in the statement equals $\phi(0) = 1$, so the claim is immediate. Suppose $S > 0$. Choose $R > 0$ such that $\nu((R,\infty)) \leq \varepsilon/2$, which is possible since ν is a probability measure. Replace each rate $r > R$ by R and let ν_R be the resulting probability measure on $[0, R]$. Since every exponential lies in $[0, 1]$ on this domain,

$$\sup_{s\in[0,S]} \left| \int e^{-rs}\nu(dr) - \int e^{-rs}\nu_R(dr) \right| \leq \nu((R,\infty)) \leq \varepsilon/2. \tag{9}$$

Partition $[0, R]$ into finitely many disjoint intervals I_k of length at most $\delta = \varepsilon/(2S)$, with endpoints assigned to one interval only. Use the right endpoint $r_k > 0$ of each interval as its representative, and set $\beta_k = \nu_R(I_k)$. Intervals of zero mass can be discarded. The mean-value theorem gives $|e^{-rs} - e^{-r_k s}| \leq S|r - r_k| \leq S\delta$ for $r \in I_k$ and $s \in [0, S]$. Hence

$$\sup_{s\in[0,S]} \left| \int e^{-rs}\nu_R(dr) - \sum_k \beta_k e^{-r_k s} \right| \leq S\delta \sum_k \beta_k = \varepsilon/2. \tag{10}$$

Combining the bounds proves the result. The coefficients are non-negative and sum to one by construction, and the representatives are positive even when ν has an atom at zero. □

Scope and normalisation. The assumption that ϕ has a positive Laplace-mixture representation is essential; mere monotonicity is insufficient. Such ϕ is continuous on $[0, S]$ by dominated convergence. For a fixed finite history and a mixture $F(s) = \sum_k \beta_k e^{-r_k s}$, write $Z_k^{\text{norm}} = \sum_{i=1}^n e^{-r_k s_i}$ and $D = \sum_k \beta_k Z_k^{\text{norm}} > 0$. The choice

$$\alpha_k = \frac{\beta_k Z_k^{\text{norm}}}{D} \tag{11}$$

gives

$$\sum_k \alpha_k \frac{e^{-r_k s_i}}{Z_k^{\text{norm}}} = \frac{F(s_i)}{\sum_j F(s_j)}. \tag{12}$$

This is an algebraic correspondence between normalised mixtures and normalised heads, not an expressivity theorem for the particular gate network. No approximation rate in K or statistical recovery claim is asserted.

D Proof of Proposition 3

Proof. All $p_i(r)$ are strictly positive for finite r. If two valid lags differ, choose i, j with $s_i \neq s_j$. Equality $\mathbf{p}(r) = \mathbf{p}(r')$ implies

$$-r(s_i - s_j) = \log \frac{p_i(r)}{p_j(r)} = \log \frac{p_i(r')}{p_j(r')} = -r'(s_i - s_j), \tag{13}$$

so $r = r'$. Next, with $\bar{s}(r) = \sum_i p_i(r) s_i$, differentiation yields $p_i'(r) = p_i(r)[\bar{s}(r) - s_i]$. Therefore

$$\bar{s}'(r) = \bar{s}(r)^2 - \sum_i p_i(r) s_i^2 = -\operatorname{Var}_{\mathbf{p}(r)}(s). \tag{14}$$

The variance is strictly positive when at least two lags differ because all positions have positive probability. For $K \geq 2$, each JSD term is non-negative and vanishes exactly when its two distributions coincide. Their sum is zero precisely when all head distributions coincide, which by injectivity is equivalent to all effective rates coinciding. If instead $s_i = s_0$ for every i, the common factor e^{-rs_0} cancels and $p_i(r) = 1/n$ for every r, proving the degenerate case. □

Implications for the regularisers. Because $\mathcal{L}_{\text{div}}$ is the negative of the divergence sum, it is always non-positive. For histories with distinct lags, it equals zero under complete effective-rate collapse and is negative if at least one pair differs. For $K > 2$, some heads can still coincide while the total divergence is positive. The finite penalty does not impose pairwise distinctness or guarantee any property of an optimisation trajectory. The load-balance penalty favours balanced mean gate usage; the simplex constraint itself comes from the softmax. Neither penalty establishes identification of a generative mixing measure. In particular, the simultaneous transformation $\lambda_k \mapsto c_k \lambda_k$ and $\tau_k \mapsto c_k \tau_k$ with $c_k \geq 1$ preserves all effective rates, attention distributions, and predictions, including under the user-conditioned modulation in Eq. equation 2.

E Full Per-Dataset Ablation (HR / NDCG / MRR)

Table A1 preserves the expanded ablation summary reported for the expanded comparison. Its Full row differs from the main-text summary on Beauty, Sports, Games and Yelp; we retain both rather than replace one reference with unverified values. Consequently, cross-table differences must not be interpreted as repeated trials of an identical configuration. Within this expanded summary, the Transformer readout improves all three sparse-data metrics relative to its listed backbone, while the ML-1M backbone is stronger. Table A2 shows the corresponding reported Mamba contrasts. These are descriptive within-table comparisons; seed-level pairing and uncertainty are not established.

Table A1: Full MARS-T ablation across all benchmarks and metrics ($\times 100$). Full values are from the expanded summary and differ from Table 4; no cross-table run identity is assumed.

	Beauty			**Sports**			**Games**			**ML-1M**			**Yelp**		
Variant	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR
Full	8.95	4.46	3.10	5.28	2.45	1.59	12.34	5.67	3.62	29.30	16.38	12.43	6.50	4.25	3.56
Backbone only	8.53	4.22	2.89	4.89	2.29	1.49	11.75	5.55	3.55	30.33	17.36	13.42	5.36	3.72	3.23
− b1	8.69	4.39	3.07	5.16	2.41	1.57	12.32	5.59	3.55	29.82	17.12	13.24	6.49	4.23	3.54
− b2	8.94	4.41	3.02	5.21	2.40	1.55	12.32	5.64	3.61	29.25	16.44	12.53	6.39	4.19	3.52
− b3	8.90	4.41	3.04	5.21	2.40	1.54	12.14	5.54	3.54	29.22	16.51	12.62	6.37	4.18	3.52
− b4	8.88	4.41	3.04	5.12	2.37	1.52	12.26	5.61	3.59	28.41	15.90	12.08	6.33	4.16	3.50
Single head	8.80	4.38	3.03	5.27	2.42	1.55	12.23	5.60	3.59	27.45	15.18	11.42	6.42	4.19	3.51

F Implementation and Reproducibility

Hyperparameters. The hyperparameters of MARS-T and MARS-M, together with their default values, are listed in Table A3. Two values are dataset-dependent: dropout (0.5 on Beauty/Sports/Yelp, 0.2 on Games/ML-1M, matching the SASRec defaults in RecBole) and the number of decay heads K (4 on the four sparse benchmarks, 8 on ML-1M). All other hyperparameters are constant across datasets and architectures.

Table A2: Reported MARS-M ablation across datasets and metrics ($\times 100$). On ML-1M, Full is above the listed matched backbone; removing diversity lowers HR@10 by 1.30 points. This difference does not demonstrate head collapse or latent-rate recovery.

	Beauty			Sports			Games			ML-1M			Yelp		
Variant	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR	HR	NDCG	MRR
Full	8.01	4.08	2.88	4.68	2.27	1.53	12.37	5.81	3.81	32.80	18.80	14.52	5.99	3.89	3.25
Backbone only	8.15	4.19	2.97	4.39	2.22	1.55	12.07	5.83	3.93	32.40	18.50	14.30	5.44	3.66	3.12
− b1	7.47	4.08	3.04	4.80	2.29	1.52	11.87	5.56	3.64	32.30	18.40	14.20	6.07	3.86	3.18
− b2	7.74	4.02	2.88	4.66	2.26	1.52	12.28	5.78	3.80	32.10	18.40	14.10	5.91	3.86	3.24
− b3	8.01	4.06	2.85	4.69	2.26	1.52	12.28	5.78	3.80	31.50	18.10	13.90	5.97	3.88	3.23
− b4	8.01	4.08	2.88	4.68	2.27	1.53	12.34	5.79	3.79	32.40	18.40	14.20	5.99	3.87	3.22
Single head	7.87	4.03	2.85	4.81	2.31	1.54	12.36	5.78	3.79	32.50	18.40	14.10	6.12	3.93	3.26

Table A3: MARS hyperparameter defaults; dataset-dependent values noted in the rightmost column.

Hyperparameter	**Default**	**Notes**
hidden dim d	64	inherited from SASRec / Mamba4Rec
#blocks N	2	Transformer blocks for MARS-T, Mamba blocks for MARS-M
inner dim	256	SASRec FFN
#decay heads K	4	8 on ML-1M
modulation strength σ	0.5	Eq. equation 2
λ_{div}	0.01	JSD diversity weight
λ_{bal}	0.01	MoE-style load balance weight
time unit τ_0	1 hour	timestamp normaliser
agg. hidden dim	64	gate MLP
gate temperature τ_α	1.0	Eq. equation 5
dropout	0.5	0.2 on Games/ML-1M (SASRec defaults)
optimiser	Adam (Kingma & Ba, 2015)	default β, no weight decay
learning rate	10^{-3}	–
batch size	2048	–
max epochs	200	with early stopping
early-stop patience	20	on validation NDCG@10

Data preprocessing. We follow the SIGMA (Liu et al., 2025) preprocessing protocol. The three Amazon datasets and ML-1M are filtered with the standard 5-core procedure (drop users and items with fewer than five interactions). Yelp uses the $[5, 200)$ filter recommended by SIGMA. Sequences are sorted by timestamp and padded or truncated to a maximum length of 50 on the four sparse datasets and 200 on ML-1M. Test items follow the leave-one-out protocol on the most recent interaction.

Code release. The intended release must identify the code revision, preprocessing and split files, configuration, seed, and checkpoint underlying each reported table. The manuscript's current aggregate summaries do not establish these correspondences. An accessible repository with matched prediction records is necessary to reproduce the reported comparisons.

G Subgroup Analysis Pseudo-code and Bucket Definitions

Bucket boundaries. The subgroup definitions are reproduced below. A distinction is necessary between raw observed-history length and the truncated input length: ML-1M inputs are capped at 200, so a ≥ 201 bucket can only refer to pre-truncation history. The available summaries do not establish which length field generated the reported cells or provide complete bucket counts. They are therefore retained as diagnostics, not used as a verified decomposition of the main comparison.

Pseudo-code. Algorithm A2 specifies the consistent evaluation procedure needed to compare aggregate and subgroup results. It uses the same checkpoint, target, and candidate mask as aggregate evaluation and retains raw length separately from truncated model input. This is a specification, not evidence that the reported subgroup summaries were generated by precisely this procedure. Full-sort model scores alone do not determine the candidate set.

Table A4: Bucket boundaries used in the subgroup analysis (Section 4.6). ML-1M uses upward-shifted boundaries to match its dense long-history regime.

Dataset	Short	Medium	Long	Very long
Beauty / Sports / Games / Yelp	1–5	6–15	16–49	≥ 50
ML-1M	1–30	31–90	91–200	≥ 201

Algorithm A2 Subgroup evaluation.

Require: trained model f, test loader $\mathcal{D}$ with raw history lengths and aggregate-evaluation candidate masks, bucket boundaries $\{(\ell_b, u_b, n_b)\}_{b=1}^{B}$, top-$k = 10$.

Initialise $H_b \leftarrow 0$, $N_b \leftarrow 0$, $M_b \leftarrow 0$, $C_b \leftarrow 0$ for each bucket b.
for batch $\mathbf{x}$, L_u^{raw}, $v^\star$ in $\mathcal{D}$ **do**
  $\mathbf{s} \leftarrow f(\mathbf{x}) \in \mathbb{R}^{B \times |\mathcal{V}|}$ ▷ full-sort scores
  Apply the same candidate mask as aggregate evaluation, including padding exclusion.
  Compute per-row rank r of $v^\star$ in $\mathbf{s}$.
  for each user u in batch **do**
    $b \leftarrow$ bucket s.t. $\ell_b \leq L_u^{\text{raw}} \leq u_b$
    $H_b \mathrel{+}= \mathbb{I}[r_u \leq k]$; $N_b \mathrel{+}= \mathbb{I}[r_u \leq k] \cdot \frac{1}{\log_2(r_u+1)}$; $M_b \mathrel{+}= \mathbb{I}[r_u \leq k] \cdot \frac{1}{r_u}$; $C_b \mathrel{+}= 1$
  end for
end for
return $\{(b, H_b/C_b,\ N_b/C_b,\ M_b/C_b)\}_{b=1}^{B}$

H Attention Entropy of the Trained SASRec Encoder

Table A5 reports normalised post-softmax attention-entropy summaries from the SASRec encoder. Entropy measures dispersion over attended positions. A lower value indicates concentration, but does not locate that concentration in the recent or distant past. Interpretation across sequence lengths also depends on the valid causal support used in the normalisation. The table is descriptive and does not establish that entropy causes the readout gains.

Table A5: Normalised post-softmax attention entropy of a trained SASRec across datasets.

Dataset	Beauty	Sports	Yelp	Games	ML-1M
Layer 0	0.785	0.812	0.838	0.870	0.512
Layer 1	0.745	0.770	0.798	0.831	0.432
Reported HR@10 contrast	$+9.95\%$	$+13.5\%$	$+19.2\%$	$+36.2\%$	$-2.9\%^\dagger$

† MARS-T vs SASRec on the reported very-long ML-1M bucket; MARS-M gains $+8.3\%$ over Mamba4Rec on the full ML-1M.

Among the sparse datasets, the reported entropy ranking matches the complete-model contrast listed here. This is not the ranking of matched-backbone increments, and the ML-1M entry is a subgroup contrast rather than an aggregate comparison. We therefore do not use this association as a mechanism test or as validation of the backbone-selection threshold.

I Hyperparameter Sensitivity Numerical Values

Table A6: Numerical values for the hyperparameter sensitivity figure (HR@10 / NDCG@10 / MRR@10, $\times 100$).

	Beauty (MARS-T, K)			**ML-1M** (K)		**Dropout**	
K / dropout	HR	NDCG	MRR	MARS-T	MARS-M	Beauty	ML-1M
1 / 0.1	8.69	4.29	2.95	27.45	32.20	8.33	31.64
2 / 0.2	8.96	4.38	2.97	27.92	32.34	8.27	30.40
4 / 0.3	9.01	4.42	3.02	28.15	32.46	8.59	27.78
8 / 0.5	9.03	4.48	3.09	28.23	32.58	8.56	25.58
16 / 0.7	8.90	4.46	3.11	29.22	32.51	8.19	22.05

J Extended Discussion

A unified interface. The common element in MARS-T and MARS-M is a timestamp-conditioned readout, not an identical encoder. This factorisation makes temporal aggregation a component that can be analysed and compared without redefining the downstream ranking interface. The observed preference for different encoders across the benchmark set supports this modular view, while leaving the selection rule empirical.

Mechanism. Effective rates provide a direct control over the temporal concentration of each head. Adaptive fusion allows the model to combine these profiles using encoded context. The theoretical results describe that control precisely, but do not identify the underlying behavioural process or ensure that trained heads remain distinct. The ablations show that increased temporal flexibility has setting-dependent value: single-head and component-reduced variants are sometimes as strong as the full model.

Limitations. The study uses offline benchmarks and only one dense dataset. The backbone threshold is not independently validated, and no comparison with CTA or MEANTIME is included in the reported experiments. Baselines have one seed, limiting uncertainty assessment. The ablation summaries use differing Full references, and the subgroup summaries do not reconcile with the aggregate table; neither supports a paired statistical interpretation without run-level provenance. Profiling differences also require a common counting scope before claiming architectural savings. These boundaries restrict the conclusions to the reported comparisons, rather than universal advantages of real-time multi-head aggregation.

Broader impact. Recency-sensitive recommendation can respond to changes in user activity, but may also amplify transient behaviour and reinforce feedback loops. Offline accuracy does not establish exposure diversity, fairness or deployment benefit. We study public benchmarks and do not assess these downstream effects.