FREE ENERGY MIXER

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard attention stores keys/values losslessly but reads them via a per-head convex average, blocking channel-wise selection. We propose the Free Energy Mixer (FEM): a free-energy (log-sum-exp) read that applies a value-driven, per-channel log-linear tilt to a fast prior (e.g., from queries/keys in standard attention) over indices. Unlike methods that attempt to improve and enrich the (q, k) scoring distribution, FEM treats it as a prior and yields a value-aware posterior read at unchanged complexity, smoothly moving from averaging to per-channel selection as the learnable inverse temperature increases, while still preserving parallelism and the original asymptotic complexity ($O(T^2)$ for softmax; $O(T)$ for linearizable variants). We instantiate a two-level gated FEM that is plug-and-play with standard and linear attention, linear RNNs and SSMs. It consistently outperforms strong baselines on NLP, vision, and time-series at matched parameter budgets.[1]

1 INTRODUCTION

Transformers, powered by attention mechanisms, have become the default backbone for sequence modeling across language, vision, speech, and decision making (Vaswani, 2017; Devlin et al., 2019; Radford, 2018; Brown et al., 2020; Dosovitskiy et al., 2020; Dong et al., 2018; Chen et al., 2021; Touvron et al., 2023). Their success is often linked to selective access to an ever-growing key-value cache while retaining parallel training and inference. In large language models, this selective ability, composed across multiple attention layers and residual pathways, supports long-range memory retrieval and the algorithmic behaviors associated with in-context learning (for example induction heads and pattern completion), as shown by recent empirical and mechanistic studies (Min et al., 2022; Wei et al., 2023; Xie et al., 2022; Zhang et al., 2023; Garg et al., 2022; Akyürek et al., 2023; Li et al., 2023; Dai et al., 2023; Bai et al., 2023; Olsson et al., 2022; Elhage et al., 2021).

Causal softmax attention combines strong selectivity with parallel efficiency: at each step it forms a distribution over past indices and mixes their values, while all steps can be computed in parallel. Given $(Q, K, V) \in \mathbb{R}^{T\times d}$ with rows $\boldsymbol{q}_t, \boldsymbol{k}_i, \boldsymbol{v}_i$, define masked scores $s_{t,i} = \boldsymbol{q}_t^\top \boldsymbol{k}_i/\sqrt{d}$ for $i \le t$ and $-\infty$ otherwise, and set $\alpha_{t,\cdot} = \mathrm{softmax}(s_{t,\cdot}) \in \Delta^{t-1}$. The step-$t$ read is

$$\boldsymbol{o}_t = \textstyle\sum_{i\le t} \alpha_{t,i}\, \boldsymbol{v}_i, \quad \boldsymbol{o}_t \in \mathrm{conv}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_t\},$$

and stacking all t yields $O = AV$ with $A_{t,i} = \alpha_{t,i}$, so a single matrix multiply produces all outputs.

The convex-mixture view explains efficiency: outputs are probability-weighted averages of the shared value bank, computed in one matrix multiply. Yet this also reveals a lossless-storage versus lossy-processing dilemma (Fig. 1a). The KV-cache stores full context, but the read is lossy: each head applies the same weights to all coordinates of $\boldsymbol{v}_i$, so $\boldsymbol{o}_t = \sum_{i\le t} \alpha_{t,i}\boldsymbol{v}_i$ lies in $\mathrm{conv}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_t\}$ and all channels are synchronized. As a result, even simple per-channel indexing, such as $s_t^\star = (v_{i_1,1}, \dots, v_{i_D,D})$ (e.g., coordinate-wise argmax), cannot be represented unless all chosen indices coincide. Adding more heads only creates a few synchronized groups, and deeper stacks cannot recover per-channel index identity once the first convex mixing has occurred. This limitation hinders Transformers in long-range modeling with non-sequential or irregular timestep indexing, and in tasks where channel-wise structure is critical, such as multivariate time series modeling (Tay et al., 2020; Zeng et al., 2023; Nie et al., 2022; Liu et al., 2024; Lu et al., 2025).

Most recent advances in attention aim to improve expressivity and efficiency, typically by designing richer selective distributions but still reading values through a token-separable linear combination.

[1] Code: anonymous.4open.science/r/Free-Energy-Mixer-6D5F.

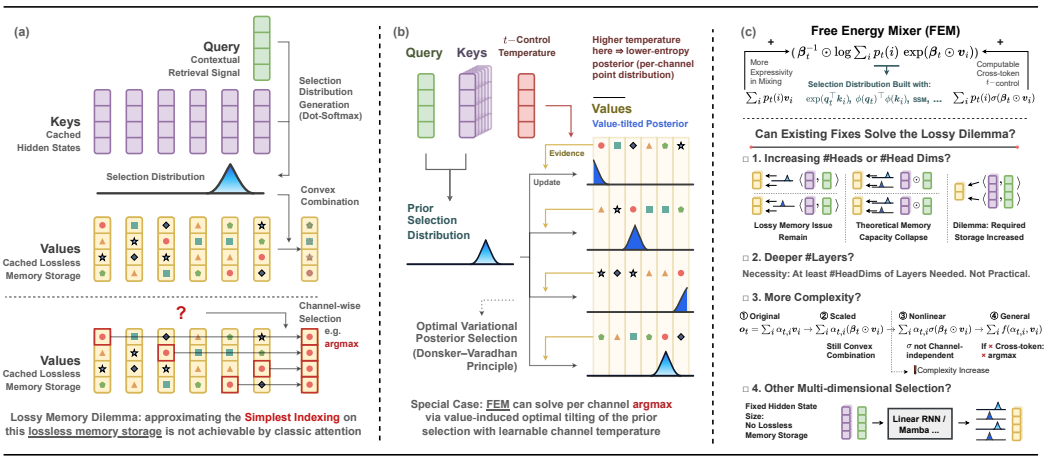

Figure 1: (a) Classic attention stores past values losslessly but reads them as a single convex combination, so channel-wise indexing (e.g., per-channel argmax) is not representable. (b) Free Energy Mixer (FEM) treats selection as a DV free-energy problem: values tilt the prior to a value-aware posterior with a learnable per-channel temperature, enabling low-entropy (point-like) posteriors and channel-wise selection while preserving the prior's time complexity. (c) Common fixes (more heads, deeper stacks, separable mixers, and per-channel scoring) either keep channels synchronized or raise cost / rely on fixed-state storage; none close the lossy-memory gap that FEM addresses.

These methods include sparsity (Beltagy et al., 2020; Child et al., 2019; Zaheer et al., 2020), low-rank projections (Wang et al., 2020; Xiong et al., 2021), and kernelizable variants with normalization or gating (Katharopoulos et al., 2020; Choromanski et al., 2021; Hua et al., 2022; Yang et al., 2024b; Qin et al., 2022a;b). Efficiency-oriented work accelerates the same semantics via factorized implementations (Dao et al., 2022; Dao, 2023) or replaces the cache with streaming state-space and RNN models of fixed size (Gu & Dao, 2023; Sun et al., 2023). Across these lines, computation is faster or the distribution richer, but the read remains a linear mix, so channels share weight vectors, and even simple per-channel indexing (e.g., argmax) cannot be realized in one step. Some recent works explore more complex combinations (e.g., nonlinear mixing such as log-sum-exp in LASER attention, or hard/top-k selection (Gupta et al., 2021; Duvvuri & Dhillon, 2025; Hashemi et al., 2025)), yet these mainly target training stability or accuracy in specific cases and do not address the lossy processing limitation .

Motivated by this gap, we propose the Free Energy Mixer (FEM), which regards lossless processing as the optimal interaction between a selection distribution and stored values: for each channel, choose an index distribution that maximizes utility under an information budget. FEM removes the linear-combination bottleneck and enables per-channel, context-dependent selection from the KV cache, while keeping causal masking, parallelism, and the asymptotic complexity of the underlying mechanism. When strong selection is not needed, FEM reduces to the standard expectation; when it is, different channels can focus on different past indices in the same step.

Contributions. (1) We identify a lossless-memory processing gap in attention: per-head convex mixing cannot realize channel-wise selection from the lossless KV-cache. (2) We propose FEM, which closes this gap by casting the read as a variational free-energy optimization that, per channel, selects an index distribution under an information budget, enabling value-aware channel-wise selection. (3) FEM is agnostic to how the selection distribution is formed (softmax, kernel/low-rank attention, linear RNNs, SSMs) and preserves the corresponding time complexity. (4) On NLP, vision, and time-series tasks, FEM consistently improves strong baselines at matched parameter sizes.

2 Methodology

2.1 Preliminaries: Selection Distributions

To analyze the storage-processing gap, we introduce the following notion of a **selection distribution**. At step t, we formalize memory selection over past indices $\mathcal{I}_t = \{1, \dots, t\}$ by a probability

vector $p_t \in \Delta^{t-1}$ with support $M_t = \{i \in \mathcal{I}_t : p_t(i) > 0\}$,

$$p_t(i) \geq 0, \qquad \sum_{i=1}^{t} p_t(i) = 1. \tag{1}$$

Causality and hard masks can be encoded by restricting M_t. Given values $\boldsymbol{v}_i \in \mathbb{R}^D$, the per-step readout is the expectation

$$o_t \;=\; \sum_{i=1}^{t} p_t(i)\, \boldsymbol{v}_i \;=\; \mathbb{E}_{i\sim p_t}[\boldsymbol{v}_i] \;\in\; \mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}. \tag{2}$$

Causal softmax self-attention is the case where p_t is a masked row-softmax over logits $\boldsymbol{q}_t^\top \boldsymbol{k}_i/\sqrt{d}$; linear attention arises when p_t is a normalized nonnegative kernel, as detailed in §B.1.

Lossless storage vs. lossy processing. Unlike RNNs, which compress history into a fixed-size state, softmax attention stores the full KV-cache $\{(\boldsymbol{k}_i, \boldsymbol{v}_i)\}_{i\leq t}$ without compression (lossless storage), but the read equation 2 applies one weight vector per head to all coordinates, so outputs lie in a per-head convex hull. This is potentially lossy when different channels should retrieve different indices in the same step. To state the target capability we define the finest-granularity retrieval:

Definition 2.1 (Channel-wise selector)**.** A channel-wise selector at time t is any vector $\boldsymbol{s}_t^\star = (v_{i_1,1}, \ldots, v_{i_D,D})$ with $i_j \in \mathcal{I}_t$ allowed to differ across $j \in [D]$.

Lemma 2.2. *Let $\boldsymbol{m}_t = (\max_{i\leq t} v_{i,1}, \ldots, \max_{i\leq t} v_{i,D})$. If $\boldsymbol{m}_t \in \mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}$, then a single index simultaneously attains all coordinate maxima. Hence if the arg-max indices differ across coordinates, $\boldsymbol{m}_t \notin \mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}$.*

Corollary 2.3. *A per-head convex read $\sum_i p_t(i)\boldsymbol{v}_i$ cannot realize a generic channel-wise selector with at least two coordinates selecting different indices.*

This geometric limitation above motivates our method. We can see that a single head applies one selection distribution to all channels at step t, synchronizing channel-wise index choices; with H heads the number of realizable head-level arg-max patterns is at most t^H, far below the t^D patterns needed for lossless per-channel selection when $H \ll D$. This gap motivates replacing the expectation read equation 2 with the free-energy read in Section 2.3. Proofs of Lemma 2.2 and Corollary 2.3, the t^H capacity counting are deferred to Appendix B.

2.2 Why standard remedies fail: toward a faithful, lossless read

We revisit common extensions around attention and explain why they do not close the channel-wise lossless-selection gap, as shown in Fig. 1c. Full details and proofs are in Appendix C.

(1) More heads. Heads provide H selection distributions per layer but synchronize channels within each head. Hence the step-t head-level argmax capacity is at most t^H, far below t^D when $H \ll D$.

Lemma 2.4. *Let $\alpha_{t,\cdot}^{(h)} \in \Delta^{t-1}$ be the distribution of head $h \in [H]$. Across contexts, realized head-level argmax assignments are at most t^H, and all coordinates controlled by head h share $\alpha_{t,\cdot}^{(h)}$.*

Increasing H reduces the per-head width $d_h = D/H$, tightening the low-rank bottleneck on the value path; as H approaches D, the cache become well-approximated by a finite-state linearization, effectively breaking the lossless-memory advantage. See Appendix C.1 for details and analysis.

(2) More depth. After a first per-head convex mixing acts at step t, per-channel index identities are no longer available unless a fresh, independent selection acts before that first mixing.

Lemma 2.5. *The map $\{\boldsymbol{v}_i\}_{i\leq t} \mapsto \sum_i \alpha_{t,i}\boldsymbol{v}_i$ is row-stochastic with image in $\mathrm{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}$. Any channel-wise selector outside this hull cannot be realized at step t by composing coordinate-wise maps and later attentions that only access already mixed tokens.*

Proposition 2.6 (Selection budget)**.** *With L attention-MLP blocks and H heads per block, at most HL disjoint channel groups receive independent first-mixing distributions by step t. A necessary condition for D independent per-channel selections at step t is $HL \geq D$ (which is not practical).*

(3) Per-dimension queries/keys. Giving each coordinate its own scoring subspace raises capacity toward t^D but raises score parameters and compute from $\Theta(d^2)$ to $\Theta(Dd)$ per layer, typically harming value bandwidth or MLP width under fixed budgets.

(4) Richer in-head mixers. The progressive family below still keeps mixing token-separable:

$$\boldsymbol{o}_t = \textstyle\sum_i \alpha_{t,i}\boldsymbol{v}_i \;\Rightarrow\; \sum_i \alpha_{t,i}(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i) \;\Rightarrow\; \sum_i \alpha_{t,i}\,\sigma(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i) \;\Rightarrow\; \sum_i f(\alpha_{t,i}, \boldsymbol{v}_i),$$

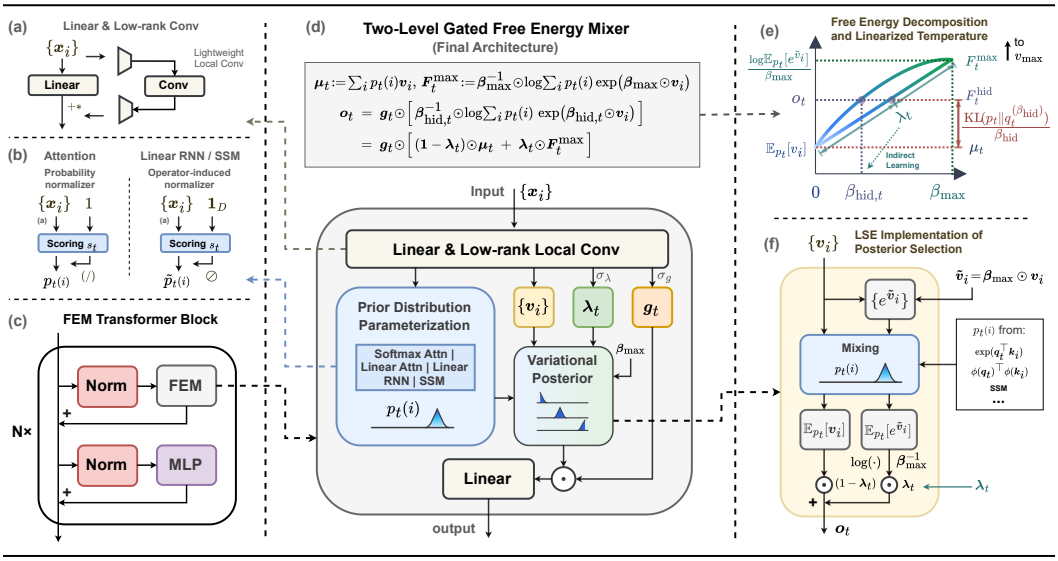

Figure 2: Overview of the Two-Level Gated Free Energy Mixer. (a) Lightweight linear & low-rank local convolution for local conditioning. (b) Prior selection: softmax attention uses a probability normalizer, while linear RNN/SSM use an operator-induced normalizer. (c) FEM integrated into a Pre-Norm Transformer block. (d) Final architecture: compute mean μ_t and max-temperature branch $F_t^{\max}$, with inner gate λ_t interpolating and outer gate g_t scaling. (e) Free-energy curve: improvement over μ_t equals $\mathrm{KL}(p_t \| q^{(\beta)})/\beta$. (f) Efficient implementation: one mixing with p_t yields both $\mathbb{E}_{p_t}[v]$ and $\beta_{\max}^{-1} \log \mathbb{E}_{p_t}[e^{\beta_{\max} v}]$, then gating produces o_t.

Proposition 2.7 (Token-separable mixers are convexly constrained)**.** *Linear and coordinate-wise gated variants lie in a convex hull of transformed values; even with a pointwise nonlinearity inside the sum, channel-wise selection of the original coordinates is not realizable in general. For general token-separable f, per-channel argmax is impossible in general. Additionally, adding per-channel cross-token competition in f may break $O(T)$/$O(T^2)$ parallelism. Details in Appendix C.4.*

(5) Linear RNNs/SSMs. They offer rich dimension interactions but store history in a fixed-size state, cannot support arbitrary index retrieval at large horizons without lossless storage; see § C.5.

Takeaway and connections. Prior remedies fall into three buckets: (a) increasing assignment capacity at substantial cost (e.g., per-feature score-space inflation to obtain $\alpha_{t,i,c}$), (b) keeping a token-separable convex read (e.g., in-head pointwise gates), or (c) relying on fixed-state storage (e.g., linear RNNs/SSMs). None provides per-channel, value-aware cross-token competition before the first mixing while preserving the time complexity. In particular, pushing capacity from t^H toward t^D via per-feature inflation leaves the read token-separable, so the same-step lossless-selection gap persists (Lemma 2.5, Proposition 2.7); likewise, simply scaling heads/depth or adding in-head gates cannot recover channel-wise index identity once the first convex mix has acted. These gaps motivate a single, stronger mixer that performs value-aware competition without changing asymptotic cost: our FEM via a variational free-energy read. See Appendix C.6 for a mapping of existing designs and Appendix C.7 for more discussion.

2.3 Free Energy Mixer: Value-Aware Posterior Selection

Motivation and objective. Classic attention performs a per-head convex read and cannot realize same-step channel-wise selectors in general (cf. Lemma 2.2, Corollary 2.3). We therefore cast channel-wise retrieval as an information-constrained selection problem: at step t, a fast, information-sparse prior p_t (from queries/keys or an operator-induced normalizer) proposes indices on the masked support M_t, while values $\{\boldsymbol{v}_i\}$ supply evidence.[2] For each channel j we choose $q \in \Delta(M_t)$

[2]Somewhat counterintuitively, we treat selection as prior and values as evidence because evidence requires log-exp processing while the prior does not; this preserves the time complexity of the selection mechanism.

to maximize expected utility under a KL budget relative to p_t,

$$\max_{q\in\Delta(M_t)} \mathbb{E}_{i\sim q}[v_{i,j}] \quad \text{s.t.} \quad \mathrm{KL}\big(q\|p_t\big) \le B_{t,j}. \tag{3}$$

Free Energy Mixer formulation. Introducing a Lagrange multiplier $\beta_{t,j} > 0$ yields the per-channel free energy output

$$\mathcal{F}_{t,j}(\beta_{t,j}) \;=\; \tfrac{1}{\beta_{t,j}} \log \textstyle\sum_{i\in M_t} p_t(i)\, \exp\big(\beta_{t,j}\, v_{i,j}\big), \tag{4}$$

and the corresponding posterior selection distribution

$$q^{(j)}_{t,\beta}(i) \;=\; \frac{p_t(i)\, \exp\big(\beta_{t,j} v_{i,j}\big)}{\sum_{r\in M_t} p_t(r)\, \exp\big(\beta_{t,j} v_{r,j}\big)}\,, \qquad i \in M_t. \tag{5}$$

Theorem 2.8 (Free-energy selection and budget duality (with β as inverse temperature))**.** *The constrained problem equation 3 has a unique solution $q^\star$. There exists a unique $\beta^\star_{t,j} \ge 0$ such that $q^\star = q^{(j)}_{t,\beta^\star}$ and $\mathbb{E}_{q^\star}[v_{i,j}] = \mathcal{F}_{t,j}(\beta^\star_{t,j})$. Equivalently (DV form), for any $\beta > 0$ the maximizer of $\sum_i q(i)v_{i,j} - \frac{1}{\beta}\mathrm{KL}(q\|p_t)$ is $q^{(j)}_{t,\beta}$. Moreover, $\beta \mapsto \mathcal{F}_{t,j}(\beta)$ is continuous and strictly increasing unless $v_{\cdot,j}$ is p_t-a.s. constant. See Appendix E, Lemmas E.1–E.2 and Proposition E.3.*

Consequences (summary). (i) Reverse-KL improvement over the mean: $\mathcal{F}_{t,j}(\beta) = \mathbb{E}_{p_t}[v_{i,j}] + \frac{1}{\beta}\mathrm{KL}(p_t\|q^{(\beta)}_{t,j})$ (Proposition E.3). (ii) Value-aware competition: the gradient equals the posterior and the Hessian is a Fisher covariance scaled by β; thus $\mathcal{F}_{t,j}$ is convex and $\beta/2$-smooth in $v_{\cdot,j}$ (Proposition E.4). (iii) Channel-wise selection on the prior support: with margin $\Delta_{t,j} > 0$, $q^{(\beta)}_{t,j}$ concentrates at the argmax with exponentially small error in β; $\mathcal{F}_{t,j}(\beta) \uparrow \max_i v_{i,j}$ (Proposition E.5). (iv) Capacity and complexity: across channels, FEM attains the assignment upper bound $|M_t|^D$, whereas H heads offer at most $|M_t|^H$ patterns; computing equation 4 with a *fixed* temperature adds one masked log-sum-exp per channel and preserves the prior's asymptotic complexity (Theorem E.7, Proposition E.8). (v) Masks and invariances: masking is preserved; shift/scale laws and sensitivity to prior probabilities/logits follow from log-sum-exp structure (Proposition E.6).

Outputs. FEM exposes two per-channel readouts sharing the same posterior $q^{(j)}_{t,\beta}$: the free energy $\mathcal{F}_{t,j}(\beta)$ and the posterior expectation $\sum_i q^{(j)}_{t,\beta}(i)\, v_{i,j}$. Under β-concentration they coincide at the selected value—letting the model smoothly move from averaging to hard indexing without changing the architecture. In § 2.3.1–2.3.2 we add a lightweight two-level gating and linearized temperature learning that learn a dynamic temperature without changing the prior's asymptotic complexity.

2.3.1 Efficient Computation of FEM and Linearized Temperature Learning

Fixed temperature. For a fixed inverse temperature $\beta > 0$ and channel j, FEM reads

$$\mathcal{F}_{t,j}(\beta) \;=\; \frac{1}{\beta} \log \sum_{i\in M_t} p_t(i)\, e^{\beta\, v_{i,j}} \;=\; \mathbb{E}_{i\sim p_t}[v_{i,j}] \;+\; \frac{1}{\beta}\, \mathrm{KL}\big(p_t \,\|\, q^{(\beta)}_{t,j}\big), \tag{6}$$

with posterior selector $q^{(\beta)}_{t,j}(i) \propto p_t(i)\, e^{\beta v_{i,j}}$ on the same support M_t as the prior. Evaluating equation 6 requires a single masked log-sum-exp (LSE) per channel, so the asymptotic time complexity is identical to the prior (e.g., $O(T^2)$ for softmax, $O(T)$ for kernel/SSM priors). See Appendix F.1.

Why β should be dynamic. The decomposition in equation 6 reveals an energy-entropy trade-off: β governs the improvement over the expectation baseline through $\frac{1}{\beta}\mathrm{KL}(p_t\|q^{(\beta)}_{t,j})$. Tasks typically need different entropy levels across steps and channels, but directly recomputing equation 6 for each learned $\beta_{t,j}$ would break single-pass efficiency.

Linearized Temperature Learning (LTL). Fix a per-channel maximum $\beta_{\max} > 0$ and define the expectation baseline $\mu_{t,j} = \mathbb{E}_{i\sim p_t}[v_{i,j}]$ and the high-temperature branch $\mathcal{F}^{\max}_{t,j} = \beta^{-1}_{\max} \log \sum_{i\in M_t} p_t(i) e^{\beta_{\max} v_{i,j}}$. A learned gate $\lambda_{t,j} \in [0,1]$ interpolates

$$\widetilde{\mathcal{F}}_{t,j}(\lambda_{t,j}) = (1-\lambda_{t,j})\, \mu_{t,j} + \lambda_{t,j}\, \mathcal{F}^{\max}_{t,j}, \tag{7}$$

requiring only the baseline expectation and a single LSE at $\beta_{\max}$ per step, hence preserving the prior's asymptotic complexity, as shown in Figure 2f.

Hidden temperature and equivalent reparameterization in LTL. Let $F_{t,j}(\beta) = \mathcal{F}_{t,j}(\beta)$ and $\Delta_{t,j}(\beta) = F_{t,j}(\beta) - F_{t,j}(0)$. Then $F'_{t,j}(\beta) = \beta^{-2}\,\mathrm{KL}\big(q^{(\beta)}_{t,j} \| p_t\big) \geq 0$, so $F_{t,j}$ is continuous and strictly increasing on $[0, \beta_{\max}]$ unless $v_{\cdot,j}$ is p_t-a.s. constant. By the intermediate value theorem, for each $\lambda_{t,j} \in [0,1]$ there exists a unique

$$\beta^\star_{t,j}(\lambda_{t,j}) = \Delta^{-1}_{t,j}(\lambda_{t,j}\,\Delta_{t,j}(\beta_{\max})) \in [0, \beta_{\max}] \quad \text{such that} \quad \widetilde{\mathcal{F}}_{t,j}(\lambda_{t,j}) = F_{t,j}\big(\beta^\star_{t,j}(\lambda_{t,j})\big). \tag{8}$$

Therefore, optimizing $\lambda_{t,j}$ is a strictly monotone reparameterization of optimizing a hidden temperature β^*_{hid} for equation 4, as shown in Figure 2e; see Proposition F.2.

Final form of FEM and complexity. Collecting terms gives the per-channel read

$$\widetilde{\mathcal{F}}_{t,j}(\lambda_{t,j}) = (1-\lambda_{t,j})\,\textstyle\sum_{i\in M_t} p_t(i)\,v_{i,j} + \frac{\lambda_{t,j}}{\beta_{\max}} \log \sum_{i\in M_t} p_t(i)\,e^{\beta_{\max} v_{i,j}}, \tag{9}$$

equal to $F_{t,j}$ at the unique hidden temperature $\beta^\star_{\mathrm{hid},t,j}(\lambda_{t,j})$. Both terms can be obtained in one pass by mixing $[\,v_{i,j},\, e^{\beta_{\max} v_{i,j}}\,]$ with the same $p_t(i)$. Hence LTL achieves dynamic temperature control without changing the prior's asymptotic complexity. A KL interpretation appear in § F.3–F.4.

2.3.2 Two-Level Gated FEM: value-aware inner gating and outer modulation

We present the two-level gated FEM that turns a prior selection distribution $p_t \in \Delta^{t-1}$ into a per-channel, value-aware read while preserving the prior's time complexity. All operations below act element-wise over channels $j \in [D]$; $\odot$ and $\oslash$ denote Hadamard product and division. Let $\boldsymbol{\beta}_{\max} \in \mathbb{R}^D_{>0}$ be a learnable global maximum inverse temperature, and let $\boldsymbol{\lambda}_t \in [0,1]^D$ and $\boldsymbol{g}_t \in \mathbb{R}^D_{>0}$ be per-channel gates at step t, parameterized from the current token features. We apply sigmoid and softplus activations, and normalize $\boldsymbol{g}_t$ with RMSNorm so that its modulation does not overly distort the norm of $\boldsymbol{o}_t$. In what follows, whenever we refer to FEM, we default to this two-level gated version. Proofs and details of this section appear in Appendix G.

Inner (temperature) gate via one-pass linearized temperature learning. Define the expectation baseline and a single high-temperature branch

$$\boldsymbol{\mu}_t = \textstyle\sum_i p_t(i)\,\boldsymbol{v}_i \in \mathbb{R}^D, \qquad \boldsymbol{F}^{\max}_t = \boldsymbol{\beta}^{-1}_{\max} \odot \log \textstyle\sum_i p_t(i) \exp\big(\boldsymbol{\beta}_{\max} \odot \boldsymbol{v}_i\big) \in \mathbb{R}^D,$$

which can be obtained in one pass by mixing $[\,\boldsymbol{v}_i,\, \exp(\boldsymbol{\beta}_{\max} \odot \boldsymbol{v}_i)\,]$ with $p_t(i)$. The inner gate as hidden temperature interpolates

$$\widetilde{\boldsymbol{F}}_t(\boldsymbol{\lambda}_t) = (\mathbf{1} - \boldsymbol{\lambda}_t) \odot \boldsymbol{\mu}_t + \boldsymbol{\lambda}_t \odot \boldsymbol{F}^{\max}_t. \tag{10}$$

Outer gate and final read. The outer gate modulates the inner read:

$$\boldsymbol{o}_t = \boldsymbol{g}_t \odot \widetilde{\boldsymbol{F}}_t(\boldsymbol{\lambda}_t) = \boldsymbol{g}_t \odot \Big[(\mathbf{1} - \boldsymbol{\lambda}_t) \odot \boldsymbol{\mu}_t + \boldsymbol{\lambda}_t \odot \boldsymbol{F}^{\max}_t\Big]. \tag{11}$$

Note that the outer gating can be regarded as applying an scaling after the token mixing in free energy with hidden temperature, i.e., $\beta^{*-1}_{\mathrm{hid},t,j} \log\Big[\sum_{i\in M_t} p_t(i)\,\exp\big(\beta^*_{\mathrm{hid},t,j}\,v_{i,j}\big)\Big]^{g_{t,j}}$. For smoother optimization, we therefore parameterize $\boldsymbol{g}_t$ as strictly positive by default. Computing equation 10–11 matches the asymptotic time complexity of the prior p_t (e.g., $O(T^2)$ for softmax; $O(T)$ for kernel/SSM priors) as shown in the section above.

Containment of common mixer families. The two-level gate subsumes several widely used mixers: (i) setting $\boldsymbol{\lambda}_t = \mathbf{0}$ yields per-channel linear reweighting $\boldsymbol{o}_t = \sum_i p_t(i)\,(\boldsymbol{g}_t \odot \boldsymbol{v}_i)$; (ii) $0 < \boldsymbol{\lambda}_t < \mathbf{1}$ gives a monotone, convex mean→real-softmax interpolation per channel, enabling value-aware thresholding; (iii) letting $\boldsymbol{\lambda}_t, \boldsymbol{g}_t$ depend on $(\mathrm{ctx}, p_t, \boldsymbol{\mu}_t, \boldsymbol{F}^{\max}_t)$ realizes a broad token-separable class $\sum_i f(\alpha_{t,i}, \boldsymbol{v}_i)$ while introducing cross-token competition through the log-sum-exp branch.

2.3.3 FEM and selection distributions: a prior-agnostic interface

FEM only requires a nonnegative, normalized selection prior $p_t \in \Delta^{t-1}$ over indices $\mathcal{I}_t = \{1, \ldots, t\}$ with masked support $M_t = \{i \leq t : p_t(i) > 0\}$, and the variational read always enforces $q_t \ll p_t$. Any streaming or parallel mechanism that produces nonnegative scores $s^+_t(i) \geq 0$ induces a valid prior via the normalization $p^+_t(i) = \frac{s^+_t(i)}{\sum_{r\leq t} s^+_t(r)}$ $(i \leq t)$.

Proposition 2.9 (Complexity-preserving normalization). *If $s_t^+(i)$ is produced by an associative operator (e.g., kernelized/linear attention, linear RNN, or SSM) that admits an $O(1)$ streaming update per step, then the denominator is obtained by applying the same operator to an all-ones stream, so forming p_t preserves the asymptotic complexity of the underlying mechanism. Under FEM with fixed or LTL-controlled temperature, the read adds one masked log-sum-exp per channel on the prior support and thus preserves $O(T^2)$ (softmax) or $O(T)$ (linear/SSM) cost (See Fig. 2b).*

Parameter budgeting. Let the input/value width be D and let FEM use working width d on the value path. We allocate a ratio $r > 0$ of parameters to the prior (e.g., Q, K and, where applicable, a decay gate). Ignoring biases/norms, the per-head linear parameters decompose as $4Dd + Ddr$, covering value, output, temperature and outer gates, and the prior block of size $D \times (rd)$. To match the classic $4D^2$ budget in standard attention: (i) $d = \frac{D}{2}$, $r = 4$ (keeps Q, K at width D); (ii) $d = \frac{2D}{3}$, $r = 2$ (balanced split). See Appendix H.7 for the split and costs. In our experiments we default to (i) since it uses a forward pass with exactly the same shape as standard attention. Notably, (i) actually reduces the dimension of the value part needed to be stored in the KV-cache by half. Subsequent experimental results show that FEM's fine-grained processing allows it to achieve superior performance over priors while using an even smaller memory state cache.

Instantiations of s_t (and p_t) We adopt the following FEM selection priors as examples. (i) Softmax attention recovers the standard masked row-softmax prior. (ii) Gated linear attention (Yang et al., 2024b) keeps an associative $O(T)$ form by combining a feature kernel with an input-conditioned decay. (iii) Linear RNNs admit nonnegative bilinear scores with normalization from the same recurrence. (iv) SSM/Mamba-style priors use nonnegative impulse responses; a channel-interactive variant lifts the index set to pairs (i, k) and normalizes per output channel, enabling cross-channel competition. All formulas, streaming recurrences, and complexity details appear in Appendix H.

Low-rank convolution. Recent sequence models such as Mamba and DeltaNet (Gu & Dao, 2023; Yang et al., 2024c;a) variants commonly enhance feature parameterization with local convolutions. We adopt this idea in FEM by inserting a lightweight adaptive low-rank convolution module that produces local, position-sensitive features. Concretely, it forms a simple time-decay kernel with $O(1)$ streaming updates, so the overall cost is only $O(TH_c)$ with the low-rank dimension $H_c \ll D$ ($H_c = d/16$ by default). The resulting features modulate both the selection prior and the FEM gates, providing local adaptivity. See § I and § K for more details.

FEM as a universal fast-weight programmer. FEM provides a unified mechanism that upgrades expectation-based reads into value-aware, per-channel posterior selection while preserving the complexity. It combines temporal mixing, entropy control, local conditioning, and dual gating, thereby serving as a general and efficient fast-weight programmer Schmidhuber (1992) detailed in § J.

3 Empirical Evaluation

We evaluate the two-level gated Free Energy Mixer (FEM) with different selection priors across synthetic, NLP, CV, and time-series tasks. Specifically, we test FEM with softmax attention (FEM-SM), gated linear attention (FEM-GLA), and on selected tasks also with Mamba (FEM-Mamba) and linear RNNs using AFT (Zhai et al., 2021) (FEM-AFT) (see § H). Unless otherwise noted, we use parameter budgeting strategy (i) from § 2.3.3, which matches the parameter size of standard attention. Under this setting, FEM reuses existing efficient implementations (e.g., FlashAttention, FlashLinearAttention) for the core prior mixer (see Fig. 2d;f) with only minor value-path overhead. Our main focus is algorithmic: exploring improved mathematical structures (see § C.7). Due to limited compute and lack of fused CUDA kernels, we scale models modestly but provide fine-grained metrics and extensive ablations to highlight FEM's advantages. For ablation, we denote FEM modules as (C: low-rank convolution, L: LSE mixing, T: linearized temperature learning, G: outer gate). For example, FEM-SM (-G,T) removes outer gating and temperature learning, equivalent to SMAttn (+C,L). Unless specified, default FEM variants include all modules (C,L,T,G). We make sure that every variants have same parameter sizes with the parameter budgeting. Aside from causal autoregressive FEM shown above, encoder-only use simply removes masking. In all experiments, FEM directly replaces the attention in a Transformer block (Fig. 2c) without altering other components (MLPs, embeddings, hyperparameters). More implementation details appear in § K; datasets in § L.

MAD. We first evaluate FEM on the synthetic MAD benchmark (Poli et al., 2024), which probes sequence models on in-context tasks. As shown in Table 1, FEM-SM outperforms all other baselines (Hyena, DeltaNet, Linear Attention, Mamba2, Gated DeltaNet, Differential Transformer, (Poli et al., 2023; Yang et al., 2024b;c; Dao & Gu, 2024; Yang et al., 2024a; Ye et al., 2025)) by a clear margin. In particular, different FEM variants show strong gains on the Compress & Recall tasks, which heavily rely on algorithmic handling of dynamic context and channel interactions. On the Compress task, FEM models achieve significant improvements over existing methods thanks to their finer-grained processing of context storage. The ablation study further reveals that the two major performance jumps over prior baselines occur after introducing +L (LSE) and +T (temperature), corroborating our earlier discussion of FEM's enhanced memory storage processing. Moreover, the ablations demonstrate that FEM can elevate linear-time methods such as GLA and Mamba (with normalized $\tilde{p}_t^+$) to a level comparable with the latest attention-based variants.

Table 1: MAD benchmark evaluation results across compression, fuzzy/in-context recall, memorization, robustness, and selective copying. **Bold** marks column best.

Model	Compress	Fuzzy Recall	In-Ctx Recall	Memorize TrainSet	Noisy Recall	Selective Copy	Avg
Hyena	44.8	14.4	99.0	89.4	98.6	93.0	73.2
DeltaNet	42.2	35.7	**99.9**	52.8	**99.9**	**99.9**	71.7
LinAttn	33.1	8.2	91.0	74.9	75.6	93.1	62.6
Mamba2	43.6	21.1	96.4	86.9	96.7	93.3	73.0
GatedDeltaNet	45.0	29.8	**99.9**	80.2	**99.9**	94.3	74.9
DiffTrans	42.9	39.0	**99.9**	83.7	97.1	95.8	76.4
FEM-SM(-G,T,L,C) (SMAttn;Transformer)	44.3	24.5	**99.9**	85.7	98.5	95.1	74.7
FEM-SM(-G,T,L) (SMAttn+C)	45.0	31.4	**99.9**	85.5	**99.9**	96.3	76.3
FEM-SM(-G,T) (SMAttn+C,L)	50.3	39.0	**99.9**	85.4	**99.9**	98.0	78.8
FEM-SM(-G) (SMAttn+C,L,T)	52.3	39.1	**99.9**	85.8	**99.9**	99.4	79.4
FEM-SM (SMAttn+C,L,T,G)	53.1	**43.1**	**99.9**	85.9	**99.9**	99.3	**80.2**
FEM-SM(-C,G,T) (SMAttn+L)	49.5	26.3	**99.9**	85.7	97.5	97.5	76.1
FEM-SM(-C,G) (SMAttn+L,T)	50.7	32.8	**99.9**	85.7	98.0	97.6	77.5
FEM-SM(-C) (SMAttn+L,T,G)	51.2	35.4	**99.9**	85.9	98.5	99.0	78.3
FEM-SM (SMAttn+C,L,T,G)	53.1	**43.1**	**99.9**	85.9	**99.9**	99.3	**80.2**
FEM-GLA(-G,T,L,C) (GLA)	40.2	8.5	91.3	81.3	86.8	76.8	64.2
FEM-GLA(-G,T,L) (GLA+C)	47.1	9.4	91.7	83.4	92.5	88.5	68.8
FEM-GLA(-G,T) (GLA+C,L, $\tilde{p}_t^+(i)$)	51.2	12.4	92.2	85.1	92.4	89.2	70.4
FEM-GLA(-G) (GLA+C,L,T, $\tilde{p}_t^+(i)$)	51.9	13.2	97.1	86.1	93.5	91.4	72.2
FEM-GLA (GLA+C,L,T,G, $\tilde{p}_t^+(i)$)	53.0	19.1	**99.9**	86.3	**99.9**	99.0	74.9
FEM-MAMBA(-G,T,L,C) (Mamba)	52.7	6.7	90.4	89.5	90.1	86.3	69.3
FEM-MAMBA(-p_t norm) (Mamba+C,L,T,G, $s_t^+(i)$)	50.5	12.8	93.4	88.9	86.3	92.2	70.7
FEM-MAMBA (Mamba+C,L,T,G, $\tilde{p}_t^+(i)$)	51.1	16.8	90.7	**89.7**	92.7	97.0	73.0
FEM-AFT(-G,T,L,C) (AFT)	50.5	9.15	63	31.1	69.2	90.1	52.2
FEM-AFT (AFT+C,L,T,G)	**55.5**	9.78	90.3	80.1	90.2	93.4	69.9

Language Modeling. We follow the experimental setup of (Yang et al., 2024a;c). Under the same training environment, we train autoregressive language models with 1.3B and 340M parameters on the FineWeb-Edu dataset (Penedo et al., 2024), using 100B and 15B sampled tokens, respectively. The models are optimized with AdamW (learning rate 4×10^{-4}, cosine annealing, 1B-token warmup), weight decay 0.1, gradient clipping of 1.0, and a batch size of 0.5M tokens. We use the LLaMA-2 tokenizer with a 32K vocabulary, and set the training context length to 4096. We adopt the Open LLM Leaderboard protocol and a suite of general-ability tasks, as shown in Tab. 2. See §L for more evaluation details.

Compared with models of the same scale, using FEM improves the overall performance of prior methods such as softmax and gated linear attention. These gains are most evident in handling longer contextual instructions, tackling more complex reasoning tasks (e.g., IFEval and ARC), and boosting accuracy across multiple QA benchmarks. This reflects FEM's ability to enhance general retrieval and context processing by extending the originally synchronized head-level prior distribution into richer channel-wise and token-wise interactions. The ablation results further confirm that introducing components like +L and +T leads to substantial performance improvements.

Table 3: Comparative analysis of image classification on ImageNet.

Model	DeiT-Tiny Top-1 Acc	DeiT-Tiny Params	DeiT-Small Top-1 Acc	DeiT-Small Params
DeiT	72.20	5.7M	79.90	22.0M
TNN	72.29	6.4M	79.20	23.4M
HGRN	74.40	6.1M	80.09	23.7M
HGRN2	75.39	6.1M	80.12	23.8M
FEM-SM	76.70	5.8M	80.45	22.3M
FEM-GLA	75.80	5.8M	80.20	22.3M

Image Modeling. We evaluate FEM on the ImageNet-1K image classification task, following Qin et al. (2024), by replacing the DeiT architecture's softmax attention with our encoder-only FEM implementation. As presented in Table 3, both FEM-SM and FEM-GLA surpass previous methods (Qin et al., 2023a;b; 2024) while maintaining parameter budgets.

Table 2: Unified language modeling evaluation results across model families and scales. Abbr: Acc_n=normalized accuracy; EM=exact match; IFE-I/P = IFEval (Inst/Prompt, strict only). Shots: MMLU-P=5, GPQA=0, BBH=3, MATH=4, MuSR=0; others 0-shot.

Model *variant*	Open LLM Leaderboard							General Ability									Ranking	
	MMLU-P (Acc↑)	GPQA (Acc$_n$↑)	BBH (Acc$_n$↑)	MATH (EM↑)	MuSR (Acc$_n$↑)	IFE-I (strict↑)	IFE-P (strict↑)	ARC-C (Acc$_n$↑)	ARC-E (Acc$_n$↑)	HS (Acc$_n$↑)	PIQA (Acc$_n$↑)	BoolQ (Acc↑)	WinoG (Acc↑)	COPA (Acc↑)	OBQA (Acc$_n$↑)	SciQ (Acc$_n$↑)	Avg Rank↓	#Top1 ↑
1.3B Params – 100B Tokens																		
DeltaNet	0.109	0.263	**0.308**	0.011	0.417	0.288	0.165	0.266	0.522	0.502	0.704	0.611	0.541	0.740	0.318	0.761	4.44	1
GSA	0.110	**0.270**	0.294	**0.013**	0.438	0.300	0.179	0.287	0.529	0.510	0.712	0.541	0.536	0.760	0.330	0.773	3.38	2
RetNet	0.110	0.252	0.293	0.001	0.384	0.056	0.024	0.271	0.489	0.480	0.701	0.583	0.533	0.710	0.324	0.736	7.63	0
HGRN	0.114	0.269	0.297	0.008	0.409	0.253	0.122	0.271	0.518	0.481	0.707	0.584	0.515	0.700	0.326	0.695	5.75	0
HGRN2	**0.115**	0.254	0.295	0.002	0.350	0.223	0.129	0.282	0.504	0.317	0.671	0.416	0.522	**0.770**	0.328	0.378	6.63	2
Transformer (SMAttn)	0.114	0.259	0.296	0.011	0.365	0.270	0.141	0.280	0.492	0.492	0.705	0.621	**0.552**	0.760	0.318	0.769	4.56	1
FEM-SM (SMAttn$_{+C,L,T,G}$)	0.113	0.262	0.303	0.012	0.451	**0.326**	**0.192**	**0.364**	**0.636**	**0.519**	**0.713**	**0.624**	0.534	0.740	**0.382**	**0.807**	**2.06**	**9**
GLA	0.114	0.259	0.295	0.006	0.427	0.272	0.157	0.277	0.482	0.488	0.702	0.574	0.541	0.690	0.326	0.721	5.63	0
FEM-GLA (GLA$_{+C,L,T,G}$)	0.112	0.258	0.297	0.009	**0.475**	0.277	0.157	0.310	0.564	0.482	0.708	0.602	0.529	0.740	0.358	0.782	3.88	1
340M Params – 15B Tokens																		
DiffTrans	0.109	0.259	0.299	0.008	0.390	0.266	0.133	0.289	0.531	0.408	0.668	0.603	**0.534**	0.690	0.330	0.734	4.38	1
GatedDeltaNet	0.113	0.260	0.296	0.010	0.421	0.258	0.133	0.276	0.527	0.396	0.662	0.588	0.527	0.710	0.338	**0.735**	4.25	1
DeltaNet	0.112	0.260	**0.300**	0.009	0.452	**0.277**	**0.150**	0.269	0.502	0.405	0.653	0.519	0.504	0.690	0.316	0.717	5.44	3
FEM-SM$_{(-G,T,L,C)}$ (SMAttn)	0.106	**0.267**	0.292	0.010	0.386	0.269	0.126	0.273	0.506	0.396	0.650	0.569	0.499	0.720	0.324	0.727	6.50	1
FEM-SM$_{(-G,T,L)}$ (SMAttn$_{+C}$)	0.113	0.254	0.296	0.009	0.388	0.246	0.122	0.277	0.507	0.403	0.664	0.583	0.515	0.670	0.320	0.728	6.63	0
FEM-SM$_{(-G,T)}$ (SMAttn$_{+C,L}$)	0.112	0.258	0.298	0.009	0.401	0.254	0.129	0.290	0.518	0.407	0.657	0.595	0.511	0.690	**0.342**	0.731	4.81	1
FEM-SM$_{(-G)}$ (SMAttn$_{+C,L,T}$)	0.112	0.261	0.297	0.010	0.421	0.266	0.144	**0.293**	0.531	**0.412**	0.668	0.593	0.519	0.710	0.338	0.716	3.31	2
FEM-SM (SM-Attn$_{+C,L,T,G}$)	0.114	0.264	**0.300**	**0.012**	0.437	0.273	0.142	0.284	**0.542**	0.409	**0.676**	**0.609**	0.523	**0.730**	**0.342**	**0.735**	**1.81**	**8**
GLA	0.110	0.258	0.289	0.007	0.415	0.228	0.109	0.247	0.478	0.366	0.637	0.547	0.489	0.640	0.294	0.649	9.38	0
FEM-GLA (GLA$_{+C,L,T,G}$)	**0.115**	0.255	0.297	0.009	**0.473**	0.241	0.123	0.271	0.493	0.397	0.644	0.592	0.510	0.680	0.331	0.683	6.56	2

Time Series Forecasting (TSF). Following Lu & Yang (2025), we evaluate FEM variants on TSF, as shown in Table 4. Across datasets, FEM surpasses both its priors and domain-specific baselines such as iTransformer (Liu et al., 2024) and PatchTST (Nie et al., 2022).

Table 4: Benchmark evaluation of TSF tasks.

Dataset	FEM SM	FEM GLA	FEM Mamba	FEM AFT	GLA	AFT	iTrans-former	Patch-TST	DLinear
Weather	0.222	0.223	0.218	0.218	0.223	0.221	0.232	0.221	0.233
Solar	0.189	0.188	0.193	0.186	0.204	0.198	0.219	0.202	0.216
ETTh1	0.419	0.418	0.421	0.414	0.418	0.421	0.454	0.413	0.422
ETTh2	0.340	0.344	0.340	0.339	0.342	0.342	0.374	0.330	0.426
ETTm1	0.341	0.345	0.346	0.344	0.357	0.351	0.373	0.346	0.347
ETTm2	0.242	0.247	0.246	0.241	0.250	0.245	0.265	0.247	0.252

Computational Cost. We evaluate the training and inference speed of FEM on a Nvidia L40S GPU. To avoid confounding factors, we use an 8-layer model with 4 heads and a hidden dimension of 512, tested on randomly generated data with a context length of 2K and a batch size of 4. As shown in Table 5, the full FEM-SM achieves comparable computational efficiency to recent model structures, even without additional engineering designs.

4 Conclusion and Limitation

We proposed the Free Energy Mixer (FEM), which reframes sequence modeling as a context-interactive selection problem to overcome the "lossless storage but lossy readout" limitation of classic attention. FEM enables value-aware, per-channel posterior selection on top of any prior (softmax/linear attention, RNNs, SSMs) and, with log-sum-exp, linearized temperature learning, and two-level gating, interpolates smoothly from averaging to near hard indexing without extra complexity. It enhances contextual fast-weight programming in theory and achieves consistent gains across NLP, vision, and time-series tasks at equal parameter budgets, with ablations highlighting LSE and temperature control as key. Overall, FEM is a plug-and-play mechanism for fine-grained context processing.

Table 5: Latency & throughput comparison (TPS in K tokens/s). Lower is better for latency; higher is better for TPS.

Model	Fwd Lat. (s)	Train Lat. (s)	Fwd TPS (K)	Train TPS (K)
GatedDeltaNet	0.016	0.042	250.4	97.8
DeltaNet	0.014	0.036	292.5	113.9
HGRN2	0.009	0.024	440.0	170.7
RWKV6	0.014	0.037	293.9	109.4
RWKV7	0.017	0.050	245.1	82.2
DiffTrans	0.018	0.041	233.3	100.6
FEM-SM (-G,T,L,C)	0.012	0.027	333.1	153.7
FEM-SM (-G,T,L)	0.015	0.033	291.5	124.6
FEM-SM (-G,T)	0.016	0.035	283.7	121.2
FEM-SM (-G)	0.017	0.040	249.7	114.6
FEM-SM	0.017	0.041	246.0	104.1

Limitation. Our work focuses on advancing the algorithmic expressivity (§C.7) rather than pursuing engineering optimizations such as custom GPU kernels or acceleration strategies. Due to limited computational resources, we were unable to scale FEM to very large models or conduct very long-context evaluations. This constrained but focused scope allowed us to highlight FEM's algorithmic contributions without heavy reliance on engineering or large-scale compute.

Ethics Statement

We evaluate FEM only on publicly available benchmarks under their licenses, without collecting personal or sensitive data. FEM's enhanced retrieval ability could be misused (e.g., surveillance or deceptive content), so responsible deployment requires privacy safeguards, bias checks, and legal compliance. We also report model sizes and training tokens, and encourage energy-aware experimentation.

Reproducibility Statement

All experiments were run under a consistent setup, with FEM modules directly replacing standard attention while keeping other components unchanged. Code, configurations, and instructions are provided in the linked repository to enable replication of our results. See the code base and §3, §K, §L for more details.

References

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=liMSqUuVg9.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019. doi: 10.48550/arXiv.1911.11641. URL https://arxiv.org/abs/1911.11641. AAAI 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ua6zuk0WRH.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. URL https://openreview.net/forum?id=fzbHRjAd8U.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European conference on computer vision*, pp. 74–92. Springer, 2022.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5884–5888. IEEE, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sai Surya Duvvuri and Inderjit S Dhillon. Laser: Attention with exponential transformation. In *Forty-second International Conference on Machine Learning*, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Guoxin Feng. Element-wise attention is all you need. *arXiv preprint arXiv:2501.05730*, 2025.

Leo Gao, Jonathan Tow, Stella Biderman, Shawn Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jasmine Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 10:8–9, 2021.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Shiwei Guo, Ziang Chen, Yupeng Ma, Yunfei Han, and Yi Wang. Scformer: Structured channel-wise transformer with cumulative historical state for multivariate time series forecasting. *arXiv preprint arXiv:2505.02655*, 2025.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. Memory-efficient transformers via top-k attention. *arXiv preprint arXiv:2106.06899*, 2021. doi: 10.48550/arXiv.2106.06899. URL https://arxiv.org/abs/2106.06899.

Baran Hashemi, Kurt Pasque, Chris Teska, and Ruriko Yoshida. Tropical attention: Neural algorithmic reasoning for combinatorial algorithms. *arXiv preprint arXiv:2505.17190*, 2025. URL https://arxiv.org/abs/2505.17190.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. URL https://arxiv.org/abs/2103.03874.

Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *International conference on machine learning*, pp. 9099–9117. PMLR, 2022.

Hugging Face. Open llm leaderboard: Scores normalization. https://huggingface.co/docs/leaderboards/en/open_llm_leaderboard/normalization, 2025. Doc. accessed 2025-09-25.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (eds.), *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 95–104. ACM, 2018. doi: 10.1145/3209978.3210006. URL https://doi.org/10.1145/3209978.3210006.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pp. 19565–19594. PMLR, 2023.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Jiecheng Lu and Shihao Yang. Linear transformers as var models: Aligning autoregressive attention mechanisms with autoregressive forecasting, 2025. URL https://arxiv.org/abs/2502.07244.

Jiecheng Lu, Yan Sun, and Shihao Yang. In-context time series predictor. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models, 2023. URL https://arxiv.org/abs/2302.10866.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. Mechanistic design and scaling of hybrid architectures, 2024. URL https://arxiv.org/abs/2403.17844.

Zhen Qin, Xiaodong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. The devil in linear transformer. *arXiv preprint arXiv:2210.10340*, 2022a.

Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022b.

Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran Zhong. Toeplitz neural network for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=IxmWsm4xrua.

Zhen Qin, Songlin Yang, and Yiran Zhong. Hierarchically gated recurrent neural network for sequence modeling. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 33202–33221. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/694be3548697e9cc8999d45e8d16fe1e-Paper-Conference.pdf.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. Hgrn2: Gated linear rnns with state expansion, 2024. URL https://arxiv.org/abs/2404.07904.

Alec Radford. Improving language understanding by generative pre-training. *OpenAI technical report*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford, CA, 2011.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399. URL https://ojs.aaai.org/index.php/AAAI/article/view/6399.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Tensorized self-attention: Efficiently modeling pairwise and global dependencies together. *arXiv preprint arXiv:1805.00912*, 2018.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*, 2023. URL https://arxiv.org/abs/2310.16049.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022. doi: 10.48550/arXiv.2210.09261. URL https://arxiv.org/abs/2210.09261.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. URL https://arxiv.org/abs/2406.01574.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14138–14148, 2021.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024a.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=ia5XvxFUJT.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024c.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OvoCm1gGhN.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer. *arXiv preprint arXiv:2105.14103*, 2021.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11106–11115. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX CONTENTS

A STATEMENT OF LLM USAGE

In this paper, LLMs were mainly used to assist with writing-related tasks, including grammar checking, wording adjustments, length reduction, layout reorganization, text formatting, formula formatting, theoretical derivation formatting, and table template generation.

We also used LLMs to search for existing methods and references in order to avoid duplicating and over-claiming. However, we did not use LLMs to conduct literature reviews, nor did LLMs replace the authors in studying the cited works. We confirm that all cited literature was read by the authors, not solely by LLMs.

During experiments, LLMs were used to assist with generating or refining experimental code and scripts, especially for bug fixing and efficiency optimization.

LLMs were not used for defining research problems, proposing ideas, designing methodologies, providing theoretical insights, or creating algorithms and model architectures.

B DETAILS AND PROOFS FOR SECTION 2.1

B.1 SELECTION DISTRIBUTIONS, SUPPORT, AND NORMALIZATION

We encode causality by restricting the feasible support to $M_t = \{1, \dots, t\}$. In softmax attention,

$$p_t(i) \;=\; \frac{\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_i/\sqrt{d})\,\mathbf{1}\{i \le t\}}{\sum_{j\le t}\exp(\boldsymbol{q}_t^\top \boldsymbol{k}_j/\sqrt{d})}.$$

In linear attention we use a nonnegative feature map $\phi : \mathbb{R}^d \to \mathbb{R}^m_{\ge 0}$ and set

$$p_t(i) \;=\; \frac{\langle \phi(\boldsymbol{q}_t), \phi(\boldsymbol{k}_i)\rangle\,\mathbf{1}\{i \le t\}}{\sum_{j\le t}\langle \phi(\boldsymbol{q}_t), \phi(\boldsymbol{k}_j)\rangle}.$$

Nonnegativity guarantees $p_t \in \Delta^{t-1}$. Row-masking is absorbed by M_t.

B.2 PROOF OF LEMMA 2.2

Let $\boldsymbol{m}_t = \sum_i \lambda_i \boldsymbol{v}_i$ with $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$. For any coordinate j, $v_{i,j} \le (\boldsymbol{m}_t)_j$ implies

$$(\boldsymbol{m}_t)_j = \sum_i \lambda_i v_{i,j} \le \sum_i \lambda_i (\boldsymbol{m}_t)_j = (\boldsymbol{m}_t)_j.$$

Thus equality holds termwise: for all i with $\lambda_i > 0$, $v_{i,j} = (\boldsymbol{m}_t)_j$. Hence every such i simultaneously attains all coordinate maxima, proving the claim.

B.3 Proof of Corollary 2.3

Let $\boldsymbol{s}_t^\star = (v_{i_1,1}, \ldots, v_{i_D,D})$ with at least two distinct indices among $\{i_j\}$. Unless the chosen $\boldsymbol{v}_{i_j}$ coincide on all selected coordinates (a measure-zero degeneracy), Lemma 2.2 implies $\boldsymbol{s}_t^\star \notin \operatorname{conv}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t\}$, so no p_t satisfies $\sum_i p_t(i)\boldsymbol{v}_i = \boldsymbol{s}_t^\star$.

B.4 Head-synchronous assignment capacity

Consider H heads at step t. Let $\alpha_{t,\cdot}^{(h)} \in \Delta^{t-1}$ be head h's selection distribution and $\iota_h = \arg\max_{i \le t} \alpha_{t,i}^{(h)}$. Channels routed through head h share the same $\alpha_{t,\cdot}^{(h)}$ at their first mixing, so the pattern is determined by $(\iota_1, \ldots, \iota_H)$ and a fixed partition of channels into heads. The number of realizable patterns is at most t^H, versus t^D for fully independent per-channel selection.

B.5 Remarks on storage versus processing

Softmax attention stores the entire set $\{(\boldsymbol{k}_i, \boldsymbol{v}_i)\}_{i \le t}$ without compression, but the per-head read equation 2 enforces one weight vector across all coordinates, which is the bottleneck for tasks requiring different indices per channel. Pointwise nonlinearities or additional depth cannot recover per-channel index identity at the same step unless a new, independent selection distribution acts before the first mixing on those channels.

C Details and proofs for Section 2.2

C.1 More heads: capacity, low-rank effects, and finite-feature linearization

Bilinear form and rank. With H heads and $d_h = D/H$,

$$\boldsymbol{y}_t = \sum_{i \le t} \Big(\sum_{h=1}^{H} \alpha_{t,i}^{(h)}\, W_O^{(h)} (W_V^{(h)})^\top \Big) \boldsymbol{x}_i = \sum_{i \le t} M_t(i)\, \boldsymbol{x}_i, \qquad \operatorname{rank}\big(W_O^{(h)} (W_V^{(h)})^\top\big) \le d_h. \quad (12)$$

Proof of Lemma 2.4. At step t, head h selects $\arg\max_i \alpha_{t,i}^{(h)}$; the Cartesian product over H heads has size at most t^H. Inside a head, all output coordinates are linear images of the same $\alpha_{t,\cdot}^{(h)}$. □

Finite-feature approximation (value-path erosion). Assuming clipped logits $|q^\top k| \le R$, a single softmax head of width d_h admits an ε-accurate finite monomial feature approximation with

$$M = \binom{N + d_h}{d_h}, \qquad N = \mathcal{O}\big(R + \log(1/\varepsilon)\big),$$

so its read is uniformly approximable by a linear streaming state of size $M \times d_v$. The full result is below.

Proposition C.1 (Dimension-dependent linearization and memory collapse for a softmax head)**.** *Consider one softmax attention head with query/key width d_h and value width d_v. Assume bounded scores and values:*

$$|q_t^\top k_i| \le R \quad (i \le t), \qquad \|v_i\|_2 \le V.$$

Fix $\varepsilon \in (0, \frac{1}{4})$ and choose $N \in \mathbb{N}$ such that

$$\sum_{n=N+1}^{\infty} \frac{R^n}{n!} \;\le\; \varepsilon.$$

Define the feature map that collects all monomials up to total degree N,

$$\phi_{N,d_h}(x) \;:=\; \Big(\tfrac{x^\alpha}{\sqrt{\alpha!}}\Big)_{|\alpha| \le N} \in \mathbb{R}^M, \qquad M \;=\; \sum_{n=0}^{N} \binom{n + d_h - 1}{d_h - 1} \;=\; \binom{N + d_h}{d_h}.$$

Then, uniformly on $\{|q^\top k| \le R\}$,

$$\left| e^{q^\top k} - \phi_{N,d_h}(q)^\top \phi_{N,d_h}(k) \right| \;\le\; \varepsilon. \tag{13}$$

Define the streaming sufficient statistics

$$S_t \;:=\; \sum_{i\le t} \phi_{N,d_h}(k_i)\, v_i^\top \in \mathbb{R}^{M\times d_v}, \qquad Z_t \;:=\; \sum_{i\le t} \phi_{N,d_h}(k_i) \in \mathbb{R}^M,$$

and the linearized readout

$$\widetilde{o}_t \;:=\; \frac{\phi_{N,d_h}(q_t)^\top S_t}{\phi_{N,d_h}(q_t)^\top Z_t} \;\in\; \mathbb{R}^{d_v}.$$

If, in addition, $\varepsilon\, e^R \le \frac{1}{2}$*, then the exact softmax output* $o_t = \sum_{i\le t} \alpha_{t,i} v_i$ *with* $\alpha_{t,i} \propto e^{q_t^\top k_i}$ *satisfies the uniform (in* t*) error bound*

$$\sup_t \;\| o_t - \widetilde{o}_t \|_2 \;\le\; 4\, V\, e^R\, \varepsilon. \tag{14}$$

Consequently, a single softmax head is $\mathcal{O}(\varepsilon)$*-approximable by a linear, streaming state of size* $M \times d_v$ *plus one* M*-vector, where*

$$M \;=\; \binom{N+d_h}{d_h} \;=\; \Theta\Big(\tfrac{N^{d_h}}{d_h!}\Big), \qquad N \;=\; \Theta\Big(R + \log\tfrac{1}{\varepsilon}\Big).$$

In particular, when $d_h = 1$ *we have* $M = N + 1 = \Theta(R + \log\frac{1}{\varepsilon})$*: the head collapses to a one-dimensional kernel-RNN-like compressed memory with arbitrarily small uniform error as* $N \to \infty$*.*

Proof. Multivariate Taylor expansion of $e^{q^\top k}$ gives $e^{q^\top k} = \sum_{n=0}^{\infty} \sum_{|\alpha|=n} \frac{q^\alpha k^\alpha}{\alpha!}$. By construction of ϕ_{N,d_h}, $\phi_{N,d_h}(q)^\top \phi_{N,d_h}(k) = \sum_{n=0}^{N} \sum_{|\alpha|=n} \frac{q^\alpha k^\alpha}{\alpha!}$, so the truncation error is the scalar exponential tail evaluated at $|q^\top k| \le R$, yielding equation 13 by the choice of N.

Let $K_t(i) := e^{q_t^\top k_i},\ \widehat{K}_t(i) := \phi(q_t)^\top \phi(k_i)$. Write $N_t = \sum_i K_t(i) v_i,\ D_t = \sum_i K_t(i)$ and $\widehat{N}_t = \sum_i \widehat{K}_t(i) v_i,\ \widehat{D}_t = \sum_i \widehat{K}_t(i)$. From equation 13 and $\|v_i\|_2 \le V$,

$$\|N_t - \widehat{N}_t\|_2 \le \varepsilon \sum_{i\le t} \|v_i\|_2 \le \varepsilon V t, \qquad |D_t - \widehat{D}_t| \le \varepsilon t.$$

Since $|q_t^\top k_i| \le R$, we have $te^{-R} \le D_t \le te^R$. If $\varepsilon e^R \le \frac{1}{2}$, then $D_t - |D_t - \widehat{D}_t| \ge \frac{1}{2}\, te^{-R}$. Using the standard ratio perturbation bound,

$$\left\| \frac{N_t}{D_t} - \frac{\widehat{N}_t}{\widehat{D}_t} \right\|_2 \;\le\; \frac{\|N_t - \widehat{N}_t\|_2}{D_t - |D_t - \widehat{D}_t|} + \frac{\|N_t\|_2}{D_t} \cdot \frac{|D_t - \widehat{D}_t|}{D_t - |D_t - \widehat{D}_t|}.$$

Because $\|N_t\|_2 \le V D_t$, the RHS is at most $\frac{\varepsilon V t}{\frac{1}{2} te^{-R}} + \frac{V \cdot \varepsilon t}{\frac{1}{2} te^{-R}} = 4Ve^R\varepsilon$, which proves equation 14.

The stated complexity follows from $M = \binom{N+d_h}{d_h}$ and Stirling's approximation; for $d_h = 1$, $M = N + 1$. □

Remark. Any common scaling (e.g., $1/\sqrt{d_h}$) in dot-product attention can be absorbed into R. Position biases can likewise be included provided the total score remains bounded by R.

Numerical illustration (state size under bounded scores). We instantiate Proposition C.1 with two practically relevant score radii: a high quantile $R \simeq 5$ and an extreme upper bound $R = 10$. For target uniform kernel error ε, choose the smallest degree N with $\sum_{n>N} R^n/n! \le \varepsilon$. The resulting hidden-state size per head (in the d_h=1 collapse) is $(N+1)\, d_v = \mathcal{O}(N\, d_v)$; across all heads with total value width $D = Hd_v$ it is $\mathcal{O}(N\, D)$.

Minimal degrees N (exact tail test).

	$\varepsilon = 10^{-4}$	$\varepsilon = 10^{-6}$	$\varepsilon = 10^{-8}$
$R = 5$	$N = 19$	$N = 22$	$N = 25$
$R = 10$	$N = 33$	$N = 36$	$N = 40$

These values satisfy the safety condition of Proposition C.1 ($\varepsilon e^R \le \frac{1}{2}$); e.g., $\varepsilon e^{10} \approx 2.2{\times}10^{-2}$ at $\varepsilon = 10^{-6}$.

Concrete state sizes (per head, d_h=1). For $\varepsilon = 10^{-6}$,

$$R = 5: \quad M = N{+}1 = 23 \;\Rightarrow\; \text{state} = (N{+}1)\,d_v = 23\,d_v \quad \text{and} \quad \mathcal{O}(N\,D) = \mathcal{O}(22\,D) \text{ overall},$$
$$R = 10: \quad M = N{+}1 = 37 \;\Rightarrow\; \text{state} = (N{+}1)\,d_v = 37\,d_v \quad \text{and} \quad \mathcal{O}(N\,D) = \mathcal{O}(36\,D) \text{ overall}.$$

Thus, under realistic bounded scores, a single softmax head with d_h=1 is equivalent (up to uniform error ε) to a linear streaming memory whose per-head size grows essentially linearly with R and only mildly with ε. For $d_h > 1$, the finite-feature dimension becomes $M = \binom{N+d_h}{d_h} = \Theta(N^{d_h}/d_h!)$, explaining the strong dependence on per-head width.

C.2 Depth: no same-step unmixing and selection budget

Proof of Lemma 2.5. $\mathcal{T}_\alpha : \{\boldsymbol{v}_i\} \mapsto \sum_i \alpha_{t,i}\boldsymbol{v}_i$ is linear, nonnegative, and weight-summing to 1, hence images lie in $\mathrm{conv}\{\boldsymbol{v}_i\}$. Composing coordinate-wise maps keeps outputs in a convex hull of transformed points and does not reveal per-channel indices used before mixing. Later attentions at step t operate on a finite set of already mixed tokens; a selector outside $\mathrm{conv}\{\boldsymbol{v}_i\}$ is unreachable without a fresh independent selection before the first mixing touching those coordinates. □

Proof of Proposition 2.6. Define a channel group as coordinates whose first attention-based mixing shares the same head at some layer. Across L layers there are at most HL groups. Each group gets one independent selection distribution for its first mixing, hence at most HL independent per-channel selections by step t. Necessity of $HL \ge D$ follows; achieving the bound requires avoiding re-synchronization before first attention. □

Accumulation. Layer ℓ writes $\boldsymbol{V}^{(\ell)} \in \mathbb{R}^{t\times D}$ to KV. Stored channels scale as LD, independently selectable groups as LH; the fraction of non-independently-selectable channels does not vanish unless H scales with D.

C.3 Per-dimension queries/keys: capacity–budget tradeoff

Giving each coordinate j its own scoring subspace increases assignment capacity toward t^D, but increases parameters and compute from $\Theta(d^2)$ to $\Theta(Dd)$ per layer. Under a fixed budget this forces shrinking D (hurting value bandwidth) or the MLP width (hurting global capacity), both detrimental in long-context regimes.

C.4 Token-separable mixers remain convexly constrained

We analyze

$$\boldsymbol{o}_t = \sum_i \alpha_{t,i}\boldsymbol{v}_i \Rightarrow \sum_i \alpha_{t,i}(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i) \Rightarrow \sum_i \alpha_{t,i}\,\sigma(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i) \Rightarrow \sum_i f(\alpha_{t,i}, \boldsymbol{v}_i),$$

with coordinate-wise σ.

Proposition C.2 (Full statement of Proposition 2.7)**.** *(i) The first two are linear; images lie in* $\mathrm{conv}\{\boldsymbol{v}_i\}$ *and* $\mathrm{conv}\{\boldsymbol{\beta}_t \odot \boldsymbol{v}_i\}$ *up to coordinate-wise scaling. (ii) For* $\sum_i \alpha_{t,i}\sigma(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i)$*, outputs lie in* $\mathrm{conv}\{\sigma(\boldsymbol{\beta}_t \odot \boldsymbol{v}_i)\}$*; recovering a channel-wise selector of the original coordinates is impossible in general unless special degeneracies (e.g., identical selected coordinates across candidates) hold. (iii) For a general token-separable* f*, per-channel argmax over original coordinates is impossible in general.*

Proof sketch. (i) Direct. (ii) If $\boldsymbol{m} = (\max_i v_{i,1}, \dots)$ is outside $\mathrm{conv}\{\boldsymbol{v}_i\}$ (Lemma 2.2), any convex combination of transformed values cannot map back to $\boldsymbol{m}$ unless σ is globally invertible and aligned simultaneously across all candidates, which fails generically. (iii) Duplication argument in $D = 1$: take two identical tokens u at indices $i \ne j$ but target $\max$ to prefer one index; any token-separable $\sum_i f(\alpha_{t,i}, v_i)$ is invariant under swapping the two, contradicting index-sensitive selection. □

Complexity remark. Per-channel cross-token operations (e.g., top-k, per-channel log-sum-exp) introduce non-separable normalizations over t and typically break fused $O(T)$/$O(T^2)$ implementations.

C.5 Linear RNNs and SSMs lack lossless storage

Let $\boldsymbol{h}_t \in \mathbb{R}^S$ be a fixed-size state updated by a (possibly input-dependent) contractive linear operator. Classical lower bounds for linear time-invariant systems imply existence of sequences and horizons t where single-token recovery error from $\boldsymbol{h}_t$ is bounded away from 0 for any fixed S. Hence fixed-state models cannot provide lossless storage of $\{\boldsymbol{v}_i\}_{i\le t}$ for arbitrary index retrieval at step t, in contrast to a KV cache, and thus cannot realize channel-wise selection over all past values.

C.6 Connections to recent per-channel variants

The families in Section 2.2 subsume many contemporary designs:

(i) Score-space inflation per feature. Tensorized/multi-dimensional attention and element-wise attention allocate a scoring subspace per coordinate to produce $\alpha_{t,i,c}$ (Shen et al., 2018; Feng, 2025). This moves assignment capacity from t^H toward t^D, but the read stays token-separable, hence subject to the convex-hull constraint (Proposition C.2). Moreover, the per-feature distributions are typically prior-only (value-agnostic) at the same step, so no value-aware cross-token competition is introduced before first mixing (cf. Lemma 2.5). The parameter/compute cost also scales from $\Theta(d^2)$ to $\Theta(Dd)$ per layer; see Appendix C.3.

(ii) More heads/depth. Increasing H adds only H independent selection groups, bounding head-level assignments by t^H (Lemma 2.4); depth increases storage but not the number of independent first-mixing distributions per step beyond HL (Proposition 2.6). Hence the channel-wise lossless-selection gap remains unless H scales with D.

(iii) In-head pointwise gating. Adding coordinate-wise gates inside the per-head mixer keeps token separability (the form $\sum_i f(\alpha_{t,i}, \boldsymbol{v}_i)$), so outputs remain in a convex hull of transformed values and cannot realize per-channel argmax of the original coordinates in general (Proposition C.2). Making the gates index-sensitive requires cross-token competition per channel, which naively breaks $O(T)$/$O(T^2)$ implementations; see Appendix C.4.

Summary. Across (i)–(iii), either capacity increases at significant parameter/compute cost while the read remains token-separable, or the same convex bottleneck persists, or fixed-state storage limits retrieval. None provides per-channel, value-aware cross-token competition before the first mixing under the prior's asymptotic complexity.

C.7 Why a stronger algorithmic mixing structure matters

A mixer that natively performs value-aware, per-channel cross-token competition at the first mixing step has two practical advantages under fixed budgets:

Separation of roles. The mixer shoulders dynamic fast-weight programming (context-dependent routing/selection), while MLPs focus on feature synthesis and knowledge consolidation. In a kernel/NTK view, this corresponds to adapting the effective kernel online at the mixing site, reducing the burden on downstream static nonlinearities.

Parallelism and efficiency. If such competition is realized without changing the asymptotic complexity of the selection prior (e.g., by computing a per-channel log-partition over the same masked support), we preserve the $O(T^2)$ softmax or $O(T)$ streaming behavior and fused-kernel practicality. This is the design objective satisfied by FEM in the next subsection: it introduces value-aware, per-channel posterior selection via a variational free-energy read while retaining the prior's time complexity.

D ADDITIONAL DISCUSSION: PER-CHANNEL SCORE DISTRIBUTIONS VS. TOKEN-SEPARABLE MIXERS

Many recent variants extend a single per-head distribution $p_t = \alpha_{t,\cdot}$ to per-channel distributions $Q_t(c,\cdot) = \alpha_{t,\cdot,c} \in \Delta^{t-1}$, yielding

$$o_{t,c} \;=\; \sum_{i\le t} \alpha_{t,i,c}\, v_{i,c}, \qquad \mathbf{o}_t \;=\; \sum_{i\le t} \underbrace{\mathrm{Diag}\big(\alpha_{t,i,1},\dots,\alpha_{t,i,D}\big)}_{=:D_{t,i}}\, \phi(\mathbf{v}_i) \;=\; \sum_{i\le t} \boldsymbol{\omega}_{t,i} \odot \phi(\mathbf{v}_i), \tag{15}$$

where ϕ acts coordinate-wise and $\boldsymbol{\omega}_{t,i} = (\alpha_{t,i,1},\dots,\alpha_{t,i,D})$. Expression equation 15 is token-separable: the outer sum is over tokens and introduces no cross-token interaction inside the mixer. Consequently, for each channel c, $o_{t,c} \in \mathrm{conv}\{v_{1,c},\dots,v_{t,c}\}$, and exact coordinate-wise selection at the same step is unattainable unless $\alpha_{t,\cdot,c}$ degenerates to a point mass (cf. Lemma 2.2, Lemma 2.5, Proposition C.2).

Assignment capacity vs. convexity. Per-channel scoring lifts head-synchronous capacity from t^H to the natural upper bound t^D: across contexts, independent argmax patterns $\{i_c^\star\}_{c\in[D]}$ can in principle be realized by $\{\alpha_{t,\cdot,c}\}$ (Shen et al., 2018; Feng, 2025). However, the mixer remains a convex expectation per channel; without value-aware cross-token competition, the distributions need not concentrate on the value argmax, and the lossless-selection gap remains.

Mapping of representative designs.

- **Per-dimension score inflation.** Tensorized/multi-dimensional and element-wise attentions instantiate $\alpha_{t,i,c}$ by combining a shared token-to-token term with per-channel terms or by per-channel distances (Shen et al., 2018; Feng, 2025). These methods increase assignment capacity (toward t^D) but keep the token-separable convex read in equation 15 and are typically prior-only (depending on (q,k) but not v).
- **In-head mixer enrichments.** Per-channel rescaling, pointwise nonlinearities, or FiLM-style gates fit $\sum_i \alpha_{t,i}\,\sigma(\beta_t \odot v_i)$ and remain within Proposition C.2: the image is a convex hull of transformed values, and no same-step unmixing arises without an additional independent selection before first mixing (cf. Lemma 2.5).
- **Axis/channel attention and structural re-partition.** Methods that attend over channels (or axes) rather than over past indices change the domain of selection but do not produce per-channel distributions across time; thus they do not affect channel-wise index capacity over $\mathcal{I}_t$; see, e.g., channel-token attention in vision and time–channel layouts in forecasting (Ding et al., 2022; Liu et al., 2024; Guo et al., 2025).
- **Linear RNNs/SSMs and kernel priors.** Streaming fast-weight priors with fixed-size state offer cross-dimension couplings yet lack lossless storage over all past indices; kernelized/linearized priors preserve streaming complexity but still yield expectation reads (Katharopoulos et al., 2020; Choromanski et al., 2021; Gu & Dao, 2023).

Where FEM differs. FEM preserves the chosen prior p_t (softmax, kernel, RNN/SSM) but upgrades the read from an expectation to the free energy $\beta^{-1}\log\sum_i p_t(i)\exp(\beta v_{i,c})$, yielding per-channel, value-aware posteriors $q_{t,\beta}^{*(c)}(i) \propto p_t(i)\, e^{\beta v_{i,c}}$. This introduces cross-token competition per channel before first mixing, achieves the $|M_t|^D$ assignment capacity and admits exponential posterior concentration while retaining the prior's asymptotic time complexity (see §2.3.1).

E THEORETICAL PROPERTIES OF FEM

We fix a timestep t, a channel $j \in [D]$, the prior selection distribution p_t with support $M_t := \{i : p_t(i) > 0\}$, and the values $\{v_{i,j}\}_{i\in M_t}$.

Notation. For $\beta > 0$, define the per-channel free energy and posterior selector

$$\mathcal{F}_{t,j}(\beta) \;:=\; \frac{1}{\beta}\log\sum_{i\in M_t} p_t(i)\, e^{\beta\, v_{i,j}}, \qquad q^{(\beta)}_{t,j}(i) \;:=\; \frac{p_t(i)\, e^{\beta v_{i,j}}}{\sum_{r\in M_t} p_t(r)\, e^{\beta v_{r,j}}},\;\; i \in M_t, \tag{16}$$

and let $v_{\cdot,j} \in \mathbb{R}^{|M_t|}$ collect $\{v_{i,j}\}_{i\in M_t}$.

Standing assumptions. All statements are over the support M_t and assume $p_t(i) > 0$ for $i \in M_t$. For $\beta < \infty$, the posterior $q^{(\beta)}_{t,j}$ is unique; in the limit $\beta \to \infty$, ties may persist if margins vanish, which does not affect finite-β claims.

Lemma E.1 (Equivalence of budgeted and penalized forms)**.** *Fix t, j and a budget $B \geq 0$. The constrained problem equation 3 has a unique maximizer $q^\star \in \Delta(M_t)$. There exists a unique $\beta^\star \geq 0$ such that $q^\star = \arg\max_q\{\sum_i q(i)v_{i,j} - \frac{1}{\beta^\star}\mathrm{KL}(q\|p_t)\}$; conversely, for every $\beta \geq 0$, the maximizer of the penalized objective solves equation 3 for the budget $B = \mathrm{KL}(q^{(\beta)}\|p_t)$. The map $B \mapsto \beta^\star(B)$ is continuous and strictly increasing whenever $v_{\cdot,j}$ is not p_t-a.s. constant.*

Lemma E.2 (Donsker-Varadhan variational principle and mirror ascent)**.** *For every $\beta > 0$,*

$$\mathcal{F}_{t,j}(\beta) \;=\; \max_{q\in\Delta(M_t)} \Big\{ \sum\nolimits_i q(i)\, v_{i,j} \;-\; \tfrac{1}{\beta}\,\mathrm{KL}\big(q\|p_t\big) \Big\}, \tag{17}$$

with the unique maximizer $q^{(\beta)}_{t,j}$ in equation 16. Equivalently,

$$q^{(\beta)}_{t,j} \;=\; \arg\min_{q\in\Delta(M_t)} \; \tfrac{1}{\beta}\mathrm{KL}(q\|p_t) \;-\; \langle q,\, v_{\cdot,j}\rangle, \tag{18}$$

i.e., an exponentiated-gradient (mirror ascent) step from p_t with step β along $v_{\cdot,j}$.

Proof. Standard DV identity: $\log\sum_i p_i e^{\beta v_i} = \max_q\{\beta\langle q, v\rangle - \mathrm{KL}(q\|p)\}$. Divide by β and apply KKT; uniqueness holds on $\Delta(M_t)$ since the objective is strictly concave in q. □

Proposition E.3 (Expectation baseline and monotonicity)**.** *Let $\mu_{t,j} := \mathbb{E}_{p_t}[v_{i,j}]$. Then*

$$\mathcal{F}_{t,j}(\beta) \;=\; \mu_{t,j} \;+\; \frac{1}{\beta}\,\mathrm{KL}\big(p_t \,\|\, q^{(\beta)}_{t,j}\big) \;\geq\; \mu_{t,j}. \tag{19}$$

Moreover, $\beta \mapsto \mathcal{F}_{t,j}(\beta)$ is continuous and strictly increasing unless $v_{\cdot,j}$ is p_t-a.s. constant, with

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\,\mathcal{F}_{t,j}(\beta) = \frac{1}{\beta^2}\,\mathrm{KL}\big(q^{(\beta)}_{t,j} \,\|\, p_t\big) \;\geq\; 0,\; \mathcal{F}_{t,j}(\beta) = \mu_{t,j} + \frac{\beta}{2}\,\mathrm{Var}_{p_t}(v_{i,j}) + O(\beta^2)\,(\beta\to 0). \tag{20}$$

Proof. equation 19 follows by direct algebra using $q^{(\beta)} \propto p\, e^{\beta v}$. Differentiate $\beta^{-1}\log\sum_i p_i e^{\beta v_i}$ to obtain equation 20. The small-β expansion is the second cumulant of $v_{i,j}$ under p_t. □

Proposition E.4 (Local geometry: gradient, curvature, smoothness)**.** *$\mathcal{F}_{t,j}(\beta)$ is convex and C^∞ in $v_{\cdot,j}$, with*

$$\nabla_{v_{\cdot,j}}\mathcal{F}_{t,j}(\beta) \;=\; q^{(\beta)}_{t,j}, \qquad \nabla^2_{v_{\cdot,j}}\mathcal{F}_{t,j}(\beta) \;=\; \beta\Big(\mathrm{Diag}(q^{(\beta)}_{t,j}) - q^{(\beta)}_{t,j} {q^{(\beta)}_{t,j}}^{\top}\Big) \;\succeq\; 0. \tag{21}$$

Hence $\|\nabla\mathcal{F}_{t,j}\|_1 = \|q^{(\beta)}_{t,j}\|_1 = 1$ and $\|\nabla\mathcal{F}_{t,j}\|_2 = \|q^{(\beta)}_{t,j}\|_2 \leq 1$. Moreover, $\mathcal{F}_{t,j}$ is $\beta/2$-smooth in ℓ_2:

$$\big\|\nabla^2\mathcal{F}_{t,j}(\beta)\big\|_{\mathrm{op}} = \beta\,\lambda_{\max}\big(\mathrm{Diag}(q) - qq^\top\big) \;\leq\; \beta/2, \tag{22}$$

and the bound is tight when q is supported on two coordinates equally, e.g. $q = (1/2, 1/2, 0, \ldots, 0)$.

Proof. For equation 21, differentiate equation 16 with respect to $v_{\cdot,j}$ to obtain $\nabla\mathcal{F}_{t,j}(\beta) = q^{(\beta)}_{t,j}$ and $\nabla^2\mathcal{F}_{t,j}(\beta) = \beta\big(\mathrm{Diag}(q) - qq^\top\big)$, where $q := q^{(\beta)}_{t,j}$. Convexity and smoothness (indeed C^∞) follow from the log-sum-exp structure. The ℓ_1- and ℓ_2-norm statements follow since q is a probability vector: $\|q\|_1 = 1$ and $\|q\|_2^2 = \sum_i q_i^2 \leq \sum_i q_i = 1$.

For equation 22, write $J(q) := \mathrm{Diag}(q) - qq^\top$. This is the covariance matrix of a one-hot random vector with class-probabilities q, hence $J(q) \succeq 0$. To bound its spectral norm, apply the Gershgorin disc theorem. Row i has diagonal entry $q_i(1-q_i)$ and the sum of absolute values of the off-diagonal entries is $\sum_{j\neq i} q_i q_j = q_i(1-q_i)$, so every eigenvalue lies in

$$\bigcup_i \left[\, q_i(1-q_i) - q_i(1-q_i),\ q_i(1-q_i) + q_i(1-q_i) \,\right] = \bigcup_i \left[\, 0,\ 2q_i(1-q_i) \,\right].$$

Therefore $\lambda_{\max}\big(J(q)\big) \le \max_i 2q_i(1-q_i) \le 1/2$, with equality attained when q is supported on two coordinates equally, e.g. $q = (1/2, 1/2, 0, \ldots, 0)$ (then $J(q)$ has eigenvalues $\{0, 1/2, 0, \ldots, 0\}$). Multiplying by β gives $\|\nabla^2 \mathcal{F}_{t,j}(\beta)\|_{\mathrm{op}} = \beta\, \|J(q)\|_{\mathrm{op}} \le \beta/2$, and the bound is tight in the stated case. □

Proposition E.5 (Range, concentration, and finite-β guarantees)**.** *Let $i^\star = \arg\max_{i\in M_t} v_{i,j}$ and $\Delta_{t,j} := v_{i^\star,j} - \max_{i\neq i^\star} v_{i,j} \ge 0$. Then for all $\beta > 0$,*

$$\mu_{t,j} \le \mathcal{F}_{t,j}(\beta) \le v_{i^\star,j},\ 1 - q_{t,j}^{(\beta)}(i^\star) \le \frac{1 - p_t(i^\star)}{p_t(i^\star)}\, e^{-\beta\,\Delta_{t,j}},\ v_{i^\star,j} + \frac{1}{\beta}\log p_t(i^\star) \le \mathcal{F}_{t,j}(\beta) \le v_{i^\star,j}. \tag{23}$$

In particular, if $\Delta_{t,j} > 0$ then $q_{t,j}^{(\beta)} \Rightarrow \delta_{i^\star}$ and $\mathcal{F}_{t,j}(\beta) \uparrow v_{i^\star,j}$ exponentially as $\beta \to \infty$.

Proof. Upper bound: $\log \sum_i p_i e^{\beta v_i} \le \beta \max_i v_i$. Lower bounds: $\mathcal{F} = \mu + \frac{1}{\beta}\mathrm{KL}(p\|q^{(\beta)}) \ge \mu$ and $\sum_{i\neq i^\star} p_i e^{\beta v_i} \le (1-p^\star)e^{\beta(v^\star - \Delta)}$ give equation 23. □

Proposition E.6 (Mask preservation; shift/scale; prior sensitivity)**.** *(i) (Masking) If $p_t(i) = 0$ then $q_{t,j}^{(\beta)}(i) = 0$; restricting M_t can only decrease equation 17.*
(ii) (Shift/scale) For any $c \in \mathbb{R}$ and $a > 0$,

$$\mathcal{F}_{t,j}(\beta; v + c) = c + \mathcal{F}_{t,j}(\beta; v), \qquad \mathcal{F}_{t,j}(\beta; av) = a\,\mathcal{F}_{t,j}(a\beta; v).$$

(iii) (Prior sensitivity: probabilities) Viewing $\mathcal{F}_{t,j}$ as a function of $p \in \Delta(M_t)$,

$$\nabla_p \mathcal{F}_{t,j} = \frac{1}{\beta}\,\frac{e^{\beta v_{\cdot,j}}}{\sum_r p_r e^{\beta v_{r,j}}}, \qquad \nabla_p^2 \mathcal{F}_{t,j} = -\frac{1}{\beta}\,\frac{e^{\beta v_{\cdot,j}}\, e^{\beta v_{\cdot,j}\top}}{\big(\sum_r p_r e^{\beta v_{r,j}}\big)^2} \ \preceq\ 0,$$

so $\mathcal{F}_{t,j}$ is concave in p on the simplex.
(iv) (Prior sensitivity: logits)

- *For unnormalized weights $s_i > 0$ with $w_i = \log s_i$ and $\tilde{\mathcal{F}}(\beta; w) := \frac{1}{\beta}\log \sum_i e^{w_i + \beta v_{i,j}}$,*

$$\nabla_w \tilde{\mathcal{F}} = \frac{1}{\beta}\,\tilde{q}, \qquad \nabla_w^2 \tilde{\mathcal{F}} = \frac{1}{\beta}\big(\mathrm{Diag}(\tilde{q}) - \tilde{q}\,\tilde{q}^\top\big) \ \succeq\ 0,$$

  hence $\tilde{\mathcal{F}}$ is convex in w.

- *For normalized logits b with $p = \mathrm{softmax}(b)$, writing $J(r) := \mathrm{Diag}(r) - rr^\top$,*

$$\nabla_b \mathcal{F}_{t,j} = \frac{1}{\beta}\big(q_{t,j}^{(\beta)} - p\big), \qquad \nabla_b^2 \mathcal{F}_{t,j} = \frac{1}{\beta}\Big(J\big(q_{t,j}^{(\beta)}\big) - J(p)\Big),$$

  which is in general indefinite; thus $\mathcal{F}_{t,j}$ is a difference-of-convex function of b.

Proof. (i) is immediate from equation 16. For (ii), add c inside the exponent or reparameterize $\beta \mapsto a\beta$ to obtain the stated identities. For (iii), $\mathcal{F}(p) = \beta^{-1}\log\langle p, e^{\beta v}\rangle$ is a log of an affine function in p, hence concave; the displayed derivatives follow by direct differentiation. For (iv), both statements follow from standard properties of log-sum-exp: the unnormalized case is convex in w; composing with softmax yields a DC form with the given gradient and Hessian. □

Theorem E.7 (Channel-wise assignment capacity over the prior support)**.** *Let D be the number of channels and $a = (a_1, \ldots, a_D) \in M_t^D$. If each channel has a positive margin $\Delta_{t,j} := v_{a_j,j} - \max_{i\in M_t\setminus\{a_j\}} v_{i,j} > 0$, then there exist finite temperatures $\{\beta_{t,j}\}$ such that $\arg\max_i q_{t,j}^{(\beta_{t,j})}(i) = a_j$ for all j. Hence the set of achievable channel-index argmax patterns has cardinality $|M_t|^D$ (the natural upper bound). A single attention head, in contrast, yields at most $|M_t|$ patterns (all channels synchronized on one distribution).*

Proof sketch. Apply Proposition E.5 per channel and choose $\beta_{t,j}$ to concentrate posterior mass on a_j with any desired margin; counting patterns gives $|M_t|^D$. □

Proposition E.8 (Complexity preservation and stable backpropagation)**.** *For fixed β, computing $\mathcal{F}_{t,j}(\beta)$ requires one masked log-sum-exp over M_t and produces $q_{t,j}^{(\beta)}$ as the gradient equation 21. Therefore FEM preserves the asymptotic time complexity of the underlying prior (e.g., $O(T^2)$ or $O(T)$) while enabling numerically stable forward/backward passes using standard LSE/softmax primitives.*

Proof sketch. Convexity in q and Slater's condition yield strong duality for equation 3; KKT gives the log-linear form $q^\star \propto p_t\, e^{\beta v}$ with multiplier $\beta^\star$. The monotonicity follows from the derivative $\frac{d}{d\beta}\mathcal{F}(\beta) = \beta^{-2}\mathrm{KL}(q^{(\beta)}\|p_t)$. □

Remark. Lemmas E.1-E.8 establish that FEM is variationally optimal (DV), value-aware with explicit local geometry, monotone in temperature with variance-controlled small-β behavior, mask-preserving, concave in the prior p on the simplex, convex in unnormalized log-weights, and DC in normalized logits, capacity-optimal for channel-wise assignment over the prior support, and complexity-preserving with stable gradients.

F Details for Linearized Temperature Learning

F.1 Fixed temperature: decomposition and cost

Lemma F.1. *For fixed $\beta > 0$, the FEM read satisfies $\mathcal{F}_{t,j}(\beta) = \mu_{t,j} + \beta^{-1}\mathrm{KL}(p_t\|q_{t,j}^{(\beta)})$, where $\mu_{t,j} = \mathbb{E}_{p_t}[v_{i,j}]$ and $q_{t,j}^{(\beta)}(i) \propto p_t(i)e^{\beta v_{i,j}}$ on M_t. Evaluating $\mathcal{F}_{t,j}(\beta)$ adds one masked LSE per channel and preserves the prior's asymptotic complexity.*

Proof. Algebra from $q^{(\beta)} \propto p\, e^{\beta v}$ yields the identity; cost follows since the support is M_t. □

F.2 Monotonicity and hidden temperature

Proposition F.2. *Let $F_{t,j}(\beta) = \beta^{-1}\log\sum_{i\in M_t} p_t(i)e^{\beta v_{i,j}}$ and $\Delta_{t,j}(\beta) = F_{t,j}(\beta) - F_{t,j}(0)$. Then $F'_{t,j}(\beta) = \beta^{-2}\mathrm{KL}(q_{t,j}^{(\beta)}\|p_t) \geq 0$, with strict positivity unless $v_{\cdot,j}$ is p_t-a.s. constant. For any $\lambda \in [0,1]$, there exists a unique $\beta^\star_{t,j}(\lambda) \in [0,\beta_{\max}]$ such that $(1-\lambda)\mu_{t,j} + \lambda F_{t,j}(\beta_{\max}) = F_{t,j}(\beta^\star_{t,j}(\lambda))$. Moreover $\lambda \mapsto \beta^\star_{t,j}(\lambda)$ is continuous and strictly increasing.*

Proof. Differentiate F to obtain $F'(\beta) = \beta^{-2}\mathrm{KL}(q^{(\beta)}\|p)$. Continuity and strict monotonicity on $[0,\beta_{\max}]$ imply the claim by the intermediate value theorem; strict increase follows from strict positivity of F' in the nondegenerate case. □

Corollary F.3 (Reparameterization equivalence)**.** *For any smooth loss $\mathcal{L}$, optimizing $\lambda_{t,j}$ in $\mathcal{L}(\widetilde{\mathcal{F}}_{t,j}(\lambda_{t,j}))$ is a strictly monotone reparameterization of optimizing β in $\mathcal{L}(F_{t,j}(\beta))$: $\partial\mathcal{L}/\partial\lambda = (\partial\mathcal{L}/\partial F)\,F'(\beta^\star)\,(\partial\beta^\star/\partial\lambda)$ with $F'(\beta^\star) > 0$.*

F.3 KL-controlled interpretation of the gate

From $\mathcal{F}_{t,j}(\beta) - \mu_{t,j} = \beta^{-1}\mathrm{KL}(p_t\|q_{t,j}^{(\beta)})$ and equation 8,

$$\frac{1}{\beta^\star_{t,j}(\lambda)}\,\mathrm{KL}\big(p_t\|q_{t,j}^{(\beta^\star_{t,j}(\lambda))}\big) \;=\; \lambda\cdot\frac{1}{\beta_{\max}}\,\mathrm{KL}\big(p_t\|q_{t,j}^{(\beta_{\max})}\big),$$

so λ specifies the fraction of the achievable KL improvement realized at step t.

F.4 One-pass implementation

Form the augmented value stream $\bar{v}_{i,j} = [\, v_{i,j},\, e^{\beta_{\max} v_{i,j}} \,]$ and compute $\sum_{i \in M_t} p_t(i) \bar{v}_{i,j} = [\, \mu_{t,j},\, \sum_i p_t(i) e^{\beta_{\max} v_{i,j}} \,]$ in one pass. Then $\mathcal{F}^{\max}_{t,j} = \beta_{\max}^{-1} \log \big(\sum_i p_t(i) e^{\beta_{\max} v_{i,j}} \big)$ and equation 7–equation 9 follow.

F.5 Geometry, stability, and additional properties

For completeness we collect properties proved in Appendix E: (i) DV variational form $\mathcal{F}(\beta) = \max_q \{ \mathbb{E}_q[v] - \beta^{-1} \mathrm{KL}(q \| p) \}$ with maximizer $q^{(\beta)} \propto p\, e^{\beta v}$; (ii) gradient $\nabla_v \mathcal{F}(\beta) = q^{(\beta)}$, Hessian $\nabla_v^2 \mathcal{F}(\beta) = \beta(\mathrm{Diag}(q^{(\beta)}) - q^{(\beta)} q^{(\beta)\top})$ and $\beta/2$-smoothness; (iii) small-β expansion $\mathcal{F}(\beta) = \mu + \frac{\beta}{2} \mathrm{Var}_p(v) + O(\beta^2)$; (iv) mask preservation; shift/scale laws; concavity in p; convexity in unnormalized logits; difference-of-convex in normalized logits; (v) complexity preservation and capacity consequences when β is large.

F.6 Complexity summary

LTL requires one expectation and one masked LSE at $\beta_{\max}$ per channel, both over M_t, thus matching the prior's asymptotic complexity ($O(T^2)$ for softmax; $O(T)$ for kernel/SSM priors) while enabling dynamic temperature control in a single pass.

G Details for Two-level Gated FEM

G.1 Inner gate as hidden temperature: properties and proof

Lemma G.1 (Monotonicity and smoothness)**.** *$F_{t,j}$ is continuous on $[0, \infty)$, differentiable on $(0, \infty)$, and*

$$\frac{\mathrm{d}}{\mathrm{d}\beta} F_{t,j}(\beta) = \beta^{-2}\, \mathrm{KL}\big(q^{(\beta)}_{t,j} \,\|\, p_t\big) \geq 0, \tag{24}$$

with equality iff $v_{\cdot,j}$ is p_t-a.s. constant. Moreover $F_{t,j}$ is convex and $\beta/2$-smooth in $v_{\cdot,j}$.

Proof. Standard Donsker–Varadhan calculus yields $F_{t,j}(\beta) = \max_{q \in \Delta(M_t)} \{ \mathbb{E}_q[v_{\cdot,j}] - (1/\beta) \mathrm{KL}(q \| p_t) \}$. Envelope differentiation gives equation 24. Convexity/smoothness follow from the Fisher covariance of $q^{(\beta)}_{t,j}$. □

Proposition G.2 (Inner gate as hidden-temperature free energy)**.** *For each channel j and any $\lambda_{t,j} \in [0, 1]$ there exists a unique $\beta_{\mathrm{hid},t,j} \in [0, \beta_{\max,j}]$ such that*

$$\widetilde{F}_{t,j}(\lambda_{t,j}) = \beta^{-1}_{\mathrm{hid},t,j} \log \sum_i p_t(i) \exp\big(\beta_{\mathrm{hid},t,j}\, v_{i,j}\big).$$

Moreover, $\lambda_{t,j} \mapsto \beta_{\mathrm{hid},t,j}$ is strictly increasing unless $v_{\cdot,j}$ is p_t-a.s. constant.

Proof. Let $\Delta_{t,j}(\beta) = F_{t,j}(\beta) - F_{t,j}(0)$. By Lemma G.1, $\Delta_{t,j}$ is continuous, strictly increasing on $[0, \beta_{\max,j}]$ unless $v_{\cdot,j}$ is constant. For any $\lambda_{t,j} \in [0, 1]$, define

$$\beta_{\mathrm{hid},t,j} = \Delta^{-1}_{t,j}\big(\lambda_{t,j} \Delta_{t,j}(\beta_{\max,j})\big) \in [0, \beta_{\max,j}],$$

which is unique by strict monotonicity. Substituting yields $\widetilde{F}_{t,j}(\lambda_{t,j}) = F_{t,j}(\beta_{\mathrm{hid},t,j})$ as claimed. □

Reverse-KL improvement over the mean. For any $\beta > 0$,

$$F_{t,j}(\beta) = \mu_{t,j} + \frac{1}{\beta}\, \mathrm{KL}\big(p_t \| q^{(\beta)}_{t,j}\big), \tag{25}$$

so $\widetilde{F}_{t,j}(\lambda)$ improves over $\mu_{t,j}$ by a controlled reverse-KL term at the hidden temperature. This explains the mean→soft-max interpolation effect of the inner gate.

G.2 Complexity, single-pass computation, and streaming

Proposition G.3 (Complexity preservation)**.** *Computing equation 10–equation 11 requires one expectation and one masked log-sum-exp at $\boldsymbol{\beta}_{\max}$ per channel, hence matches the asymptotic time complexity of the prior p_t (e.g., $O(T^2)$ for softmax; $O(T)$ for kernel/SSM priors).*

Proof. Compute $\sum_i p_t(i)\,[\,\boldsymbol{v}_i,\ \exp(\boldsymbol{\beta}_{\max}\odot\boldsymbol{v}_i)\,]$ once, then split to obtain $\boldsymbol{\mu}_t$ and $\boldsymbol{F}_t^{\max}$. This requires one expectation and one masked log-sum-exp per channel on the prior support M_t, matching the prior's asymptotic time (softmax $O(T^2)$; kernel/SSM $O(T)$). The outer gate $\boldsymbol{g}_t$ is a pointwise modulation. □

Streaming compatibility. For associative priors (kernel/SSM), the normalized read is computed by the same scan used for p_t; concatenating a constant "$\mathbf{1}$" channel yields the normalizer and numerator in one pass. The LSE branch uses the same support M_t and thus preserves streaming.

Numerical stability. We use standard LSE stabilization per channel: subtract $\max_i(\beta_{\max,j}v_{i,j})$ inside the exponential and add it back after the logarithm. Gradient clipping for $\boldsymbol{\beta}_{\max}$ prevents overflow when tasks push toward hard selection.

G.3 Containment of mixer families

Proposition G.4 (Formal containment)**.** *(i) $\boldsymbol{\lambda}_t=\mathbf{0}$ gives $\boldsymbol{o}_t=\sum_i p_t(i)\,(\boldsymbol{g}_t\odot\boldsymbol{v}_i)$, matching per-channel linear reweighting. (ii) $0<\boldsymbol{\lambda}_t<\mathbf{1}$ yields a monotone, convex aggregator in each channel that interpolates between $\mu_{t,j}$ and $\max_i v_{i,j}$ as $\lambda_{t,j}$ increases. (iii) Allowing $\boldsymbol{\lambda}_t,\boldsymbol{g}_t$ to depend on $(ctx,p_t,\boldsymbol{\mu}_t,\boldsymbol{F}_t^{\max})$ realizes token-separable couplings of the form $\sum_i f(\alpha_{t,i},\boldsymbol{v}_i)$ and adds cross-token competition through the log-sum-exp term.*

Proof. Direct substitution of the choices for $\boldsymbol{\lambda}_t$ and identification of limits $\beta\to 0$ and $\beta\to\infty$ per channel. □

G.4 Capacity and hard-selection limits

Proposition G.5 (Capacity and limits on the prior support)**.** *With $\boldsymbol{\lambda}_t\approx\mathbf{1}$ and sufficiently large $\boldsymbol{\beta}_{\max}$, the per-channel posterior concentrates on its own arg-max over the prior support $M_t=\{i:p_t(i)>0\}$, so the achievable channel-index assignment capacity attains $|M_t|^D$. In the limit $\boldsymbol{\lambda}_t=\mathbf{0}$, FEM reduces to the expectation baseline (the original read of the selection prior).*

Proof. Fix channel j and let $\Delta_{t,j}=\min_{i\neq i^\star}(v_{i^\star,j}-v_{i,j})>0$ be the margin at the arg-max index $i^\star$. For any $\beta\geq\beta_0(\Delta_{t,j})$, the posterior $q_{t,j}^{(\beta)}$ places at least $1-\exp(-\beta\Delta_{t,j})$ mass on $i^\star$, and $F_{t,j}(\beta)\uparrow v_{i^\star,j}$. Across channels, with $\boldsymbol{\lambda}_t\approx\mathbf{1}$ and sufficiently large $\boldsymbol{\beta}_{\max}$, the joint posterior concentrates independently per channel over M_t, achieving $|M_t|^D$ distinct index assignments. Setting $\boldsymbol{\lambda}_t=\mathbf{0}$ recovers $\boldsymbol{\mu}_t$. □

G.5 Gradients and curvature

For channel j,

$$\frac{\partial F_{t,j}(\beta)}{\partial v_{i,j}}=q_{t,j}^{(\beta)}(i),\qquad \frac{\partial^2 F_{t,j}(\beta)}{\partial v_{i,j}\partial v_{r,j}}=\beta\Big(q_{t,j}^{(\beta)}(i)\mathbf{1}\{i=r\}-q_{t,j}^{(\beta)}(i)q_{t,j}^{(\beta)}(r)\Big).$$

Thus gradients are the posterior weights and the Hessian is a Fisher covariance scaled by β, giving stable, value-aware competition. Backprop through equation 10 is a convex combination of the mean and LSE branches with coefficients $\mathbf{1}-\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_t$.

G.6 Invariances and sensitivity to priors

For any constants $a_j>0$ and b_j, $F_{t,j}(\beta;a_jv_{i,j}+b_j)=a_jF_{t,j}(a_j\beta;v_{i,j})+b_j$. Multiplying prior probabilities by a positive scalar and renormalizing leaves $F_{t,j}$ unchanged; reweighting p_t within M_t shifts the posterior via $q_{t,j}^{(\beta)}\propto p_t\exp(\beta v_{\cdot,j})$, which is exploited by the outer and inner gates.

H PARAMETERIZATIONS OF THE PRIOR SELECTION p_t IN FEM

Unified interface. At step t, let the accessible index set be $\mathcal{I}_t = \{1, \dots, t\}$ and let a nonnegative score $s_t : \mathcal{I}_t \to \mathbb{R}_{\geq 0}$ define the prior selection by

$$p_t(i) \;=\; \frac{s_t(i)}{\sum_{r \leq t} s_t(r)}\,, \qquad i \leq t, \tag{26}$$

with $s_t(i) = 0$ for $i > t$ (causal mask). FEM then optimizes, per channel j, the DV free energy

$$\mathcal{F}_{t,j}(\beta) = \beta^{-1} \log \sum_{i \leq t} p_t(i)\, e^{\beta v_{i,j}}, \quad q^{*(j)}_{t,\beta}(i) \propto p_t(i)\, e^{\beta v_{i,j}}.$$

Below we specify s_t (hence p_t) for each prior family, along with the streaming recurrences and time complexity. Throughout, $M_t = \{i \leq t : s_t(i) > 0\}$ is the support carried into FEM (we enforce $q_t \ll p_t$).

H.1 SOFTMAX-ATTENTION PRIOR

Scores and normalization. Given masked scores $\ell_t(i) = \langle q_t, k_i \rangle + b_{t,i}$ with $\ell_t(i) = -\infty$ for $i > t$,

$$s_t(i) = \exp\{\ell_t(i)\}\,, \qquad p_t(i) = \frac{\exp\{\ell_t(i)\}}{\sum_{r \leq t} \exp\{\ell_t(r)\}}. \tag{27}$$

This is the standard row-softmax over causal scores.

Complexity. Matrix form $A = \mathrm{softmax}_{\mathrm{row}}(QK^\top + B + M_\triangle)$ yields $O(T^2)$ time and $O(T^2)$ memory (or $O(T^2)$ time, $O(T)$ KV-cache in the autoregressive setting).

H.2 GATED LINEAR ATTENTION (GLA) PRIOR

Positional encoding and positivity. We inject relative position with RoPE, then map queries/keys to the nonnegative orthant:

$$\tilde{\boldsymbol{q}}_t \;:=\; \mathrm{ReLU}\big(\mathrm{RoPE}(\boldsymbol{q}_t)\big) + \varepsilon \in \mathbb{R}^m_{\geq 0}, \qquad \tilde{\boldsymbol{k}}_i \;:=\; \mathrm{ReLU}\big(\mathrm{RoPE}(\boldsymbol{k}_i)\big) + \varepsilon \in \mathbb{R}^m_{\geq 0},$$

where $\varepsilon > 0$ is a small constant for numerical stability.

Decay gating. Let $g_\tau \leq 0$ be a learned (scalar / per-head / per-channel) gate and define the cumulative envelope

$$D_t \;:=\; \exp\Big(\sum_{\tau=1}^{t} g_\tau\Big) \quad \text{(clipped in practice)}.$$

The causal time-decay factor between index i and step t is $K_{t,i} = D_t D_i^{-1} = \exp(\sum_{\tau=i+1}^{t} g_\tau) \in (0, 1]$.

Scores and normalization. The nonnegative score and prior are

$$s_t(i) \;=\; K_{t,i}\, \big\langle \tilde{\boldsymbol{q}}_t, \tilde{\boldsymbol{k}}_i \big\rangle\, \mathbf{1}\{i \leq t\}, \qquad p_t(i) \;=\; \frac{s_t(i)}{Z_t}, \qquad Z_t \;=\; \sum_{r \leq t} s_t(r). \tag{28}$$

Equivalently, with an associative scan form,

$$Z_t \;=\; \Big\langle D_t\, \tilde{\boldsymbol{q}}_t,\, \underbrace{\sum_{r \leq t} D_r^{-1} \tilde{\boldsymbol{k}}_r}_{B_t} \Big\rangle, \qquad \sum_{i \leq t} s_t(i)\, v_i \;=\; \Big\langle D_t\, \tilde{\boldsymbol{q}}_t,\, \underbrace{\sum_{r \leq t} D_r^{-1} \big(\tilde{\boldsymbol{k}}_r \otimes v_r\big)}_{A_t} \Big\rangle. \tag{29}$$

Hence the baseline normalized read is

$$\mu_t \;=\; \frac{\langle D_t \tilde{\boldsymbol{q}}_t,\; A_t \rangle}{\langle D_t \tilde{\boldsymbol{q}}_t,\; B_t \rangle}.$$

Streaming recurrences. Both states update in $O(1)$ per step:

$$B_t = B_{t-1} + D_t^{-1}\tilde{\boldsymbol{k}}_t, \qquad A_t = A_{t-1} + D_t^{-1}\big(\tilde{\boldsymbol{k}}_t \otimes v_t\big).$$

A one-pass implementation appends a constant channel to values: $\bar{v}_t = [v_t; 1]$, $\bar{A}_t = \sum_{r\le t} D_r^{-1}(\tilde{\boldsymbol{k}}_r \otimes \bar{v}_r)$; then $[\text{num}, \text{den}] = \langle D_t\tilde{\boldsymbol{q}}_t, \bar{A}_t\rangle$ and $\mu_t = \text{num}/\text{den}$.

Complexity. GLA preserves the $O(T)$ streaming complexity (per head), with the same associative-scan cost as standard linear attention. FEM operates over the same support $M_t = \{i : p_t(i) > 0\}$ and adds one masked log-sum-exp per channel (at fixed or LTL-controlled temperature).

H.3 Linear RNN-Style Priors

(LRNN-softmax) AFT-style normalized exponential weights. Let $k_i \in \mathbb{R}^m$ be per-step logits and define

$$s_t(i) \;=\; \exp\{k_i\}\,\mathbf{1}\{i \le t\}, \qquad Z_t \;=\; \sum_{r\le t} \exp\{k_r\}\,. \tag{30}$$

Streaming recurrence:

$$S_t = S_{t-1} + e^{k_t} v_t, \quad Z_t = Z_{t-1} + e^{k_t}\,, \qquad \Rightarrow \quad \mathbb{E}_{p_t}[v_i] = S_t/Z_t.$$

(We stabilize with $k_i - \max_{r\le t} k_r$ in practice.)

(LRNN-decay) Input-conditioned exponential decay. Let $g_\tau \in \mathbb{R}_{\le 0}$ be a learned generator and define

$$s_t(i) \;=\; \exp\Big(\sum_{\tau=i+1}^{t} g_\tau\Big)\,\mathbf{1}\{i \le t\}. \tag{31}$$

With $\Gamma_t = \exp(\sum_{\tau\le t} g_\tau)$ we have $s_t(i) = \Gamma_t\,\Gamma_i^{-1}$ and the streaming form

$$C_t = C_{t-1} + \Gamma_t^{-1} v_t, \qquad \sum_{i\le t} s_t(i) v_i = \Gamma_t\, C_t, \quad Z_t = \Gamma_t \sum_{i\le t} \Gamma_i^{-1}.$$

Thus numerator and denominator share the envelope Γ_t, preserving $O(T)$ cost. (Conceptually, LRNN-decay recovers the decay portion of GLA without the dot-product features.)

Complexity. Both LRNN-softmax and LRNN-decay are $O(T)$ with $O(1)$ updates; FEM adds one masked log-sum-exp per channel.

H.4 SSM / Mamba-Style Priors

Positive impulse-response SSM. Consider a causal linear state-space operator with nonnegative impulse $H_\theta(\tau) \ge 0$:

$$(\mathcal{S}_\theta u)_t = \sum_{i\le t} H_\theta(t-i)\,u_i, \qquad H_\theta(\tau) = C_\Delta A_\Delta^{\tau-1} B_\Delta \mathbf{1}\{\tau \ge 1\} + D\mathbf{1}\{\tau = 0\},$$

where $(A_\Delta, B_\Delta, C_\Delta, D)$ are stable, nonnegative discretizations.

Scores and normalization. Set

$$s_t(i) \;=\; H_\theta(t-i)\,\mathbf{1}\{i \le t\}, \qquad Z_t \;=\; \sum_{r\le t} H_\theta(t-r) \;=\; (\mathcal{S}_\theta \mathbf{1})_t. \tag{32}$$

Both numerator and denominator come from the same scan (once with $u_i = v_i$, once with $u_i \equiv 1$), so the $O(T)$ streaming complexity is preserved. In practice we parameterize to ensure $H_\theta(\tau) \ge 0$ (e.g., softplus for (Δ, B, C, D) and negative-softplus for the diagonal generator).

H.5 Local Conditioning of the Prior

Let $c_t \in \mathbb{R}^{H_c}$ be the output of a learnable, $O(T)$ time-decay conditioner (low-rank causal convolution). We modulate the prior parameters and value-path gates by

$$\tilde{\theta}_t = \theta_t + G^{(p)}(c_t),\ \tilde{p}_t(\cdot) = p_t(\cdot\,; \tilde{\theta}_t),\ \tilde{v}_i = v_i \odot (1+\eta_t^{(v)}),\ \tilde{\lambda}_t = \lambda_t \odot (1+\eta_t^{(\lambda)}),\ \tilde{g}_t = g_t \odot (1+\eta_t^{(g)}),$$

where $G^{(p)}$ and $\eta^{(\cdot)}$ are small MLPs. This preserves the streaming/parallel cost of the chosen prior.

H.6 Support, Masking, and Complexity Summary

We always enforce $s_t(i) = 0$ for $i > t$ and for hard-masked indices, hence $M_t = \{i \le t : s_t(i) > 0\}$. FEM's per-channel variational step operates on M_t and adds exactly one masked log-sum-exp per channel (at a fixed or LTL-controlled temperature), so the asymptotic time complexity matches that of the prior: softmax $O(T^2)$, GLA/LRNN/SSM $O(T)$.

Prior family	Scores $s_t(i)$	Complexity (per head)
Softmax attention	$\exp\{\langle q_t, k_i\rangle + b_{t,i}\}$	$O(T^2 d)$
Gated linear attention (GLA)	$e^{\sum_{\tau=i+1}^{t} g_\tau}\ \langle \tilde{q}_t, \tilde{k}_i\rangle$	$O(Td)$
LRNN-softmax (AFT)	$\exp\{k_i\}$	$O(Td)$
LRNN-decay	$\exp(\sum_{\tau=i+1}^{t} g_\tau),\ g_\tau \le 0$	$O(Td)$
SSM/Mamba	$H_\theta(t-i),\ H_\theta(\cdot) \ge 0$	$O(Td)$

Remark (RoPE & positivity mapping). Any invertible positional transform $(q, k) \mapsto (Tq, Tk)$ can precede score evaluation. In our GLA prior we use RoPE followed by a ReLU $+\varepsilon$ mapping on both queries and keys to guarantee nonnegative feature vectors $(\tilde{q}_t, \tilde{k}_i) \in \mathbb{R}^m_{\ge 0}$ before decay gating and normalization.

H.7 Width and Parameter Budgeting for the Prior

Let the input/value width be D and let FEM use a working width d on the value path. We allocate a parameter ratio $r > 0$ for the prior parameterization (queries/keys and decay gate in GLA), scaled with d. Ignoring biases and norms, the per-head linear parameters decompose into five projections:

$$\underbrace{D \times d}_{\text{value}} + \underbrace{D \times d}_{\text{outer gate } \boldsymbol{g}} + \underbrace{D \times d}_{\text{temperature } \boldsymbol{\lambda}} + \underbrace{d \times D}_{\text{output}} + \underbrace{D \times (rd)}_{\text{prior (Q/K + decay)}} = 4\,Dd + Ddr.$$

The prior block $D \times (rd)$ is split among $\tilde{q}_t, \tilde{k}_i$ projections and the decay gate. To keep the total parameter count equal to classic attention ($4D^2$), two convenient choices are

$$\text{(i) } d = \tfrac{D}{2},\ r = 4 \qquad \text{or} \qquad \text{(ii) } d = \tfrac{2D}{3},\ r = 2,$$

since $4Dd + Ddr = 4D^2$ in both cases. In (i), the prior (Q/K) runs at D-dim width—identical to standard attention—while the value path uses $d = \frac{D}{2}$. In (ii), the value width increases to $d = \frac{2D}{3}$ with a balanced prior split (e.g., $\dim(Q) = \dim(K) = \frac{d}{2}$), and the remaining budget supports the decay gate. Both settings preserve the asymptotic time complexity of the chosen prior (softmax $O(T^2)$; GLA/LRNN/SSM $O(T)$).

I Low-rank Convolution: Time-Decay Conditioner (TDC)

I.1 Definition and streaming form

Given token features $\boldsymbol{x}_{1:T} \in \mathbb{R}^{T \times D}$, let $\widehat{\boldsymbol{x}}_t = \mathrm{LN}(\boldsymbol{x}_t)$ and choose a hidden width $H_c \ll D$. Define

$$\boldsymbol{s}_t = \mathrm{softplus}(\widehat{\boldsymbol{x}}_t W_f) \in \mathbb{R}^{H_c}, \quad \boldsymbol{u}_t = \widehat{\boldsymbol{x}}_t W_x \in \mathbb{R}^{H_c}, \quad \boldsymbol{a}_t = \mathrm{softplus}(\widehat{\boldsymbol{x}}_t W_s) \in \mathbb{R}^{H_c},$$

with $W_f, W_x, W_s \in \mathbb{R}^{D\times H_c}$. The positive envelope is

$$\boldsymbol{f}_t = \exp\Big(-\sum_{\tau=1}^{t} \boldsymbol{s}_\tau\Big) \in \mathbb{R}^{H_c} \quad \text{(element-wise)}.$$

The TDC output is the causal, input-conditioned separable convolution

$$\widetilde{\boldsymbol{h}}_t = \boldsymbol{f}_t \odot \sum_{i=1}^{t} \Big(\boldsymbol{u}_i \oslash \boldsymbol{f}_i\Big) = \sum_{i=1}^{t} \underbrace{\exp\Big(-\sum_{\tau=i+1}^{t} \boldsymbol{s}_\tau\Big)}_{\boldsymbol{K}_{t,i}} \odot \boldsymbol{u}_i \in \mathbb{R}^{H_c}. \tag{33}$$

A calibrated shortcut and projection produce the conditioning features:

$$\boldsymbol{h}_t = \mathrm{SiLU}\big(\mathrm{norm}(\boldsymbol{a}_t)\big) \odot \mathrm{LN}(\widetilde{\boldsymbol{h}}_t), \qquad \boldsymbol{c}_t = \boldsymbol{h}_t W_c \in \mathbb{R}^{D_c},$$

where $W_c \in \mathbb{R}^{H_c\times D_c}$ and $\mathrm{norm}(\cdot)$ rescales to unit ℓ_2 norm.

Proposition I.1 (Rank-1-in-time kernel and $O(1)$ updates)**.** *The kernel in equation 33 factors as $\boldsymbol{K}_{t,i} = \boldsymbol{f}_t \odot (\boldsymbol{f}_i)^{-1}$, i.e., rank-1 in time for each channel. Hence the convolution admits $O(1)$ streaming updates:*

$$\boldsymbol{C}_t = \boldsymbol{C}_{t-1} + \boldsymbol{u}_t \oslash \boldsymbol{f}_t, \qquad \widetilde{\boldsymbol{h}}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_t.$$

The per-sequence cost is $O(TH_c)$ and the per-step memory is $O(H_c)$.

Proof. By definition, $\boldsymbol{K}_{t,i} = \exp\big(-\sum_{\tau=i+1}^{t} \boldsymbol{s}_\tau\big) = \exp\big(-\sum_{\tau\le t} \boldsymbol{s}_\tau\big) \odot \exp\big(\sum_{\tau\le i} \boldsymbol{s}_\tau\big) = \boldsymbol{f}_t \odot (\boldsymbol{f}_i)^{-1}$. Substituting into equation 33 yields the stated streaming form. □

Stability. The softplus parameterization ensures $\boldsymbol{s}_t \ge 0$, hence $\boldsymbol{f}_t \in (0, 1]$ element-wise; this prevents exploding envelopes and ensures well-conditioned division in $\boldsymbol{u}_i/\boldsymbol{f}_i$ with standard ε stabilization.

I.2 Coupling TDC to FEM

We use disjoint slices of $\boldsymbol{c}_t$ to modulate (i) the parameterization of the prior selection $p_t(\cdot;\theta_t)$ and (ii) FEM's value-path gates:

$$\text{Prior modulation:} \quad \tilde{\theta}_t = \theta_t + \Delta\theta_t, \quad \Delta\theta_t = G^{(p)}(\boldsymbol{c}_t), \qquad \tilde{p}_t(i) := p_t\big(i;\tilde{\theta}_t\big), \tag{34a}$$

$$\text{Value gate:} \quad \tilde{\boldsymbol{v}}_i = \boldsymbol{v}_i \odot \big(1 + \boldsymbol{\eta}_t^{(v)}\big), \qquad \boldsymbol{\eta}_t^{(v)} \in [\eta^{(v)}] \subset \boldsymbol{c}_t, \tag{34b}$$

$$\text{Outer gate:} \quad \tilde{\boldsymbol{g}}_t = \boldsymbol{g}_t \odot \big(1 + \boldsymbol{\eta}_t^{(g)}\big), \qquad \boldsymbol{\eta}_t^{(g)} \in [\eta^{(g)}] \subset \boldsymbol{c}_t, \tag{34c}$$

$$\text{Temperature gate:} \quad \tilde{\boldsymbol{\lambda}}_t = \boldsymbol{\lambda}_t \odot \big(1 + \boldsymbol{\eta}_t^{(\lambda)}\big), \qquad \boldsymbol{\eta}_t^{(\lambda)} \in [\eta^{(\lambda)}] \subset \boldsymbol{c}_t. \tag{34d}$$

FEM then applies equation 11 with $(p_t, \boldsymbol{v}_i, \boldsymbol{g}_t, \boldsymbol{\lambda}_t)$ replaced by $(\tilde{p}_t, \tilde{\boldsymbol{v}}_i, \tilde{\boldsymbol{g}}_t, \tilde{\boldsymbol{\lambda}}_t)$, yielding position-aware, locally conditioned selection without changing the prior's asymptotic complexity.

I.3 Complexity and compatibility

Proposition I.2 (Complexity preservation)**.** *TDC adds $O(TH_c)$ time and $O(H_c)$ memory per layer and does not alter the asymptotic complexity of FEM's read, which remains $O(T^2)$ for softmax priors and $O(T)$ for kernel/SSM priors. The per-step coupling in equation 34 is pointwise in t and thus streaming-compatible.*

Relation to recent convolutional/SSM designs. TDC follows the spirit of low-rank, input-conditioned time-decay filters used in SSM/DeltaNet-style models and the local convolutional augmentations commonly paired with Mamba-like architectures. Our use is FiLM-like: TDC learns a compact context c_t that modulates both the selection prior and FEM gates, providing local adaptivity while preserving streaming costs.

Implementation notes. We apply standard LSE stabilization per channel in FEM's log-sum-exp branch, and clamp the envelope by computing $\boldsymbol{f}_t = \exp(-\mathrm{cumsum}(\boldsymbol{s}_t))$ in log-space with an ε floor. The projections $G^{(p)}$, $[\eta^{(v)}]$, $[\eta^{(g)}]$, $[\eta^{(\lambda)}]$ are small MLPs with per-channel outputs; their widths are tuned so that $H_c \ll D$.

J FEM AS A UNIVERSAL FAST-WEIGHT PROGRAMMER

Putting the pieces together, the final Free Energy Mixer realizes a unified, parallel fast-weight program:

$$\boldsymbol{o}_t \;=\; \tilde{\boldsymbol{g}}_t \odot \Big[\big(\mathbf{1} - \tilde{\boldsymbol{\lambda}}_t\big) \odot \underbrace{\mathbb{E}_{i\sim\tilde{p}_t}[\tilde{\boldsymbol{v}}_i]}_{\text{mean (high-entropy)}} \;+\; \tilde{\boldsymbol{\lambda}}_t \odot \underbrace{\boldsymbol{\beta}_{\max}^{-1} \odot \log \sum_{i\le t} \tilde{p}_t(i)\, \exp\big(\boldsymbol{\beta}_{\max} \odot \tilde{\boldsymbol{v}}_i\big)}_{\text{max free energy (low-entropy)}} \Big], \quad (35)$$

where the prior $\tilde{p}_t$ and the value-path gates $(\tilde{\boldsymbol{v}}_i, \tilde{\boldsymbol{g}}_t, \tilde{\boldsymbol{\lambda}}_t)$ are locally conditioned by TDC as in equation 34. Equation equation 35 shows that the mixer is simultaneously:

- a **temporal mixer** (log-sum-exp across indices, with causal masking and per-channel competition);
- an **entropy mixer** (inner temperature via $\tilde{\boldsymbol{\lambda}}_t$; mean↔soft-max interpolation);
- a **local-feature mixer** (position-aware modulation injected by TDC);
- a **dual-gated mixer** (inner temperature gate over indices i; outer amplitude gate over timesteps t).

Crucially, the assignment capacity over the prior support attains the upper bound $|M_t|^D$ (per-channel posterior selection), the variational objective is solved exactly (DV optimality), and the overall time complexity matches that of the chosen prior (softmax $O(T^2)$, kernel/SSM $O(T)$), up to the $O(TH_c)$ convolution overhead. FEM thus serves as a broadly applicable, universal fast-weight programmer that upgrades expectation-based reads to value-aware, memory processing without sacrificing parallel efficiency.

J.1 RELATION TO PRIOR POOLING AND SELECTION METHODS

Our Free Energy Mixer (FEM) is related to but distinct from several existing approaches:

- **Log-Sum-Exp (LSE) pooling.** FEM is not simply a generalized mean that interpolates between average and max pooling. Instead, from a Donsker–Varadhan variational view, it uses *values* to tilt an arbitrary prior distribution p_t. This yields per-channel, value-aware posteriors rather than only adjusting the softness of pooling.
- **Entmax / Sparsemax.** These operate directly on the scoring distribution over (q, k), changing how probability mass is allocated. FEM instead treats this distribution as a *prior* and introduces cross-token competition through the values. The two directions are complementary and could be combined.
- **Gumbel-Softmax / Top-k.** Such methods emphasize hard selection, sampling, or ranking, often requiring non-parallel sampling or offline sorting. In contrast, FEM remains fully differentiable, parallel in one pass, and preserves the asymptotic complexity of the underlying prior.

K ADDITIONAL IMPLEMENTATION DETAILS

Our detailed experimental setup is available in the linked code repository. All language modeling experiments, including both training and inference, were conducted on 8× Nvidia H100 GPUs, while all non-language modeling tasks were trained on 8× Nvidia L40S GPUs. We use 42 as the random seed. The training and inference precision is bfloat16. For each task, we replaced the standard Transformer block with an FEM Transformer block, substituting the attention layer with

FEM-{SM, GLA, Mamba, AFT}, while keeping all other settings unchanged to ensure a fully consistent experimental environment. Parameter budgeting was carefully applied to keep overall model size and architecture comparable to the baselines. The additional low-rank convolution used in our parameterization introduces less than 1% extra parameters (with $H_c = d/16$).

special configurations required by the experimental setup when specified. Otherwise, all linear projections are randomly initialized from a centered normal distribution with a standard deviation of 0.02. All biases and embeddings are initialized to zero. For the maximum inverse temperature, we initialize it to zero and then apply the parameterization softplus(x+1.8) to ensure that its initial value is around 1 and remains strictly positive throughout training.

L Additional Dataset Description

Language Model Evaluation Setup. We adopt the Open LLM Leaderboard (OLL) protocol and a complementary suite of general-ability tasks. The Open LLM Leaderboard core covers MMLU-Pro (5-shot, accuracy), GPQA (0-shot, normalized accuracy), BBH (3-shot, normalized accuracy), MATH (4-shot, exact match), and MuSR (0-shot, normalized accuracy), plus IFEval for instruction following, where we report strict pass rates for instruction- and prompt-level constraints (Wang et al., 2024; Rein et al., 2023; Suzgun et al., 2022; Hendrycks et al., 2021; Sprague et al., 2023; Zhou et al., 2023). Following OLL, we use the normalized-accuracy metric acc_n for multiple-choice tasks, which subtracts the random-guess baseline and rescales scores to a common range for fair cross-task comparison (Hugging Face, 2025). To broaden coverage, we also evaluate on widely used general-ability benchmarks: ARC (Challenge/Easy), HellaSwag, PIQA, BoolQ, WinoGrande, COPA, OpenBookQA, and SciQ, reporting accuracy or acc_n as standard; unless noted, these are evaluated in 0-shot (Clark et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Clark et al., 2019; Sakaguchi et al., 2020; Roemmele et al., 2011; Welbl et al., 2017). We perform the evaluations with lm-evaluation-harness (Gao et al., 2021).

MAD We assess our architecture using the Mechanistic Architecture Design (MAD) framework, a recently introduced methodology for cost-efficient evaluation of deep learning models Poli et al. (2024). MAD provides a set of capability-focused benchmarks—including in-context recall, fuzzy recall, selective copying, and compression—that probe core sequence modeling abilities. It has been validated across more than 500 language models ranging from 70M to 7B parameters, showing a strong correlation between performance on these synthetic tasks and compute-optimal perplexity at scale. By leveraging MAD as a reliable predictor of large-scale behavior, we can identify architectural advantages without relying on the prohibitive compute costs of full-scale training.

Time Series Forecasting We evaluate our module on several standard time series forecasting benchmarks, following the setup of Lu & Yang (2025). **(1) Weather** (Wu et al., 2021)[3]: 21 meteorological variables (e.g., temperature, humidity) collected every 10 minutes in 2020 from a German weather station. **(2) Solar** (Lai et al., 2018)[4]: Solar power output recorded every 10 minutes in 2006 from 137 U.S. photovoltaic plants. **(3) ETT** (Zhou et al., 2021)[5]: Transformer load and temperature data from July 2016 to July 2018, sampled at 15-minute (ETTm1/ETTm2) and hourly (ETTh1/ETTh2) intervals, covering 7 key operational features.

[3] https://www.bgc-jena.mpg.de/wetter/

[4] http://www.nrel.gov/grid/solar-power-data.html

[5] https://github.com/zhouhaoyi/ETDataset