Free Energy Mixer

Jiecheng Lu, Shihao Yang

Georgia Institute of Technology

Problem & Motivation

- **Lossless KV, Lossy Read.** Standard attention stores key/value pairs losslessly but reads them as a *per-head convex average*:

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

  All channels of a head share the same weights $\alpha_{t,\cdot}$, blocking channel-wise selection (e.g., per-dimension argmax).
- **Consequences.** Even with many heads/layers, once a first convex mixing happens, *per-channel index identities are lost*; capacity tops out at $\mathcal{O}(t^H)$ patterns (for H heads) vs. the t^D patterns needed for D channels.
- **Goal.** Preserve causal masking and parallelism while allowing *value-aware, channel-wise* selection from the KV cache *at the same asymptotic time complexity* as the prior (softmax or linearizable variants).

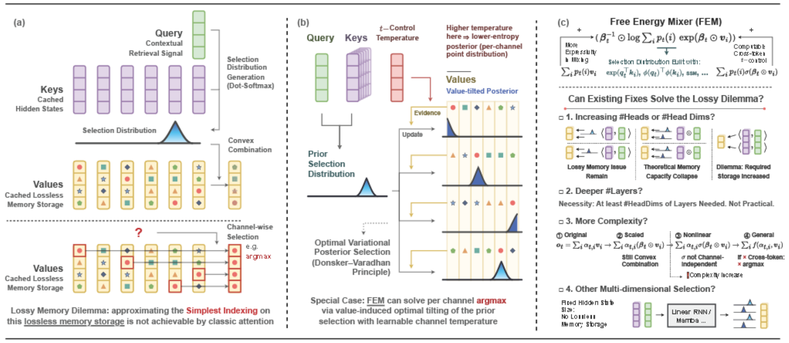

Figure 1. Lossless storage vs. lossy processing; FEM forms a value-aware posterior per channel while preserving prior complexity.

Free Energy Mixer (FEM): Core Idea

Treat the attention distribution as a **selection prior** $p_t \in \Delta^{t-1}$ over masked indices M_t, and use values as evidence to form a **value-aware posterior** *per channel*.

$$\text{Free energy read:} \quad \mathcal{F}_{t,j}(\beta) = \tfrac{1}{\beta} \log \sum_{i \in M_t} p_t(i)\, e^{\beta\, v_{i,j}},$$

$$\text{Posterior selector:} \quad q^{(j)}_{t,\beta}(i) = \frac{p_t(i) e^{\beta v_{i,j}}}{\sum_{r \in M_t} p_t(r) e^{\beta v_{r,j}}}.$$

Mean-improvement identity:

$$\mathcal{F}_{t,j}(\beta) = \mathbb{E}_{p_t}[v_{i,j}] + \frac{1}{\beta}\, \mathrm{KL}\big(p_t \,\|\, q^{(j)}_{t,\beta}\big).$$

As $\beta \uparrow$, $\mathcal{F}_{t,j}$ approaches $\max_{i \in M_t} v_{i,j}$ and $q^{(j)}_{t,\beta}$ concentrates on the maximizing index.

Linearized Temperature Learning (LTL)

Compute in one pass:

$$\mu_{t,j} = \mathbb{E}_{p_t}[v_{i,j}], \qquad \mathcal{F}^{\max}_{t,j} = \tfrac{1}{\beta_{\max}} \log \sum_{i \in M_t} p_t(i) e^{\beta_{\max} v_{i,j}}.$$

Interpolate with a learned gate $\lambda_{t,j} \in [0,1]$:

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

This is equal to $\mathcal{F}_{t,j}(\beta^\star_{t,j})$ for a *unique hidden temperature* $\beta^\star_{t,j} \in [0, \beta_{\max}]$; thus LTL learns temperature *without extra passes*.

Two-Level Gated FEM (Per Channel)

Let $\boldsymbol{\lambda}_t \in [0,1]^D$ (inner gate), $\boldsymbol{g}_t > 0$ (outer gate), and $\boldsymbol{\beta}_{\max} > 0$.

$$\boldsymbol{\mu}_t = \sum_i p_t(i) \boldsymbol{v}_i, \quad \boldsymbol{F}^{\max}_t = \boldsymbol{\beta}^{-1}_{\max} \odot \log \sum_i p_t(i) \exp\big(\boldsymbol{\beta}_{\max} \odot \boldsymbol{v}_i\big).$$

$$\widetilde{\boldsymbol{F}}_t(\boldsymbol{\lambda}_t) = (\mathbf{1} - \boldsymbol{\lambda}_t) \odot \boldsymbol{\mu}_t + \boldsymbol{\lambda}_t \odot \boldsymbol{F}^{\max}_t, \qquad \boldsymbol{o}_t = \boldsymbol{g}_t \odot \widetilde{\boldsymbol{F}}_t(\boldsymbol{\lambda}_t).$$

Special cases: (i) $\boldsymbol{\lambda}_t = \mathbf{0}$ gives per-channel linear reweighting; (ii) $\boldsymbol{\lambda}_t \in (0,1)$ yields a monotone mean→real-softmax interpolation; (iii) gating on context enables rich token-separable mixers with cross-token competition via LSE.

Key Contributions

1. **Diagnose the gap:** per-head convex mixing cannot realize generic channel-wise selection from a lossless KV cache.
2. **FEM read:** a log-sum-exp free-energy objective that yields a *value-aware posterior* per channel while preserving the prior's time complexity ($O(T^2)$ softmax, $O(T)$ linearizable).
3. **Linearized Temperature Learning (LTL):** learn *dynamic* inverse temperature in one pass by interpolating between the expectation baseline and a single high-temperature branch.
4. **Two-level gated FEM:** a plug-and-play mixer for softmax/linear attention, linear RNNs, and SSMs; consistently improves strong baselines in NLP, vision, and time series at matched parameter budgets.

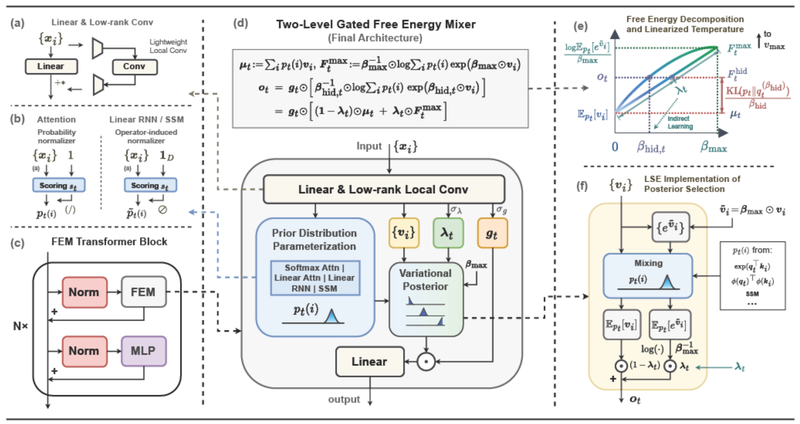

Figure 2. Two-level gated FEM: (a) local conditioning, (b) prior selection, (c) FEM in a Pre-Norm Transformer block, (d) mean & high-β branches with inner/outer gates, (e) free-energy curve, (f) one-pass implementation.

Experimental Highlights (matched parameter budgets)

Language Modeling (Open LLM Leaderboard, 1.3B). FEM-SM improves *average rank* to **2.06** with **9** top-1's, outperforming strong attention and linear-time baselines.

Table 1. Language Modeling Summary (Open LLM suite; lower AvgRank is better).

1.3B Params — 100B Tokens

Model	AvgRank	Top1
Transformer (SMAttn)	4.56	1
GLA	5.63	0
FEM-GLA	**3.88**	1
FEM-SM	**2.06**	**9**

340M Params — 15B Tokens

Model	AvgRank	Top1
SMAttn (baseline)	6.50	1
GLA	9.38	0
FEM-GLA	**6.56**	2
FEM-SM	**1.81**	**8**

Setup: FineWeb-Edu; Open LLM Leaderboard protocol. All models trained/evaluated at matched parameter budgets; FEM is a drop-in replacement for attention/prior mixers.

Benchmarks. Synthetic in-context (MAD), language modeling (1.3B/340M; FineWeb-Edu), ImageNet-1K (DeiT-Tiny/Small), and multivariate TSF (ETT, Weather, Solar). FEM is a drop-in replacement for attention/prior mixers.

MAD (Avg score, ↑). FEM-SM achieves the best average across tasks probing compression, selective copying, and noisy recall.

Model	Compress	Fuzzy	InCtx	Noisy	Select	TrainMem	Avg
Hyena	44.8	14.4	99.0	98.6	93.0	89.4	73.2
DeltaNet	42.2	35.7	**99.9**	**99.9**	**99.9**	52.8	71.7
DiffTrans	42.9	39.0	**99.9**	97.1	95.8	83.7	76.4
FEM-SM	**53.1**	**43.1**	**99.9**	**99.9**	99.3	85.9	**80.2**

Asymptotic cost preserved. FEM requires the same parallel matrix multiply (softmax) or streaming updates (linear/SSM) as the prior, plus a *single* masked log-sum-exp per channel.

Takeaways

- **What FEM gives:** A plug-and-play, prior-agnostic mixer that upgrades expectation-based reads into *value-aware, per-channel* posterior selection—bridging the gap between lossless storage and faithful readout.

Reference

[1] Lu, Jiecheng, and Shihao Yang. "Free Energy Mixer." In Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026), 2026.