IMPROVING BAYESIAN IN-CONTEXT LEARNING IN MARKOV TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

A transformer can enter a generalizing in-context learning regime while remaining far from the corresponding optimal predictor. What prevents it from closing this residual gap? We study this question in Markov-chain prediction, where the Bayes posterior is available exactly and the learned attention circuit can be inspected. We introduce *Symmetry-Aware In-context Pre-training* (SAIP) for two-block transformers. SAIP guides the formation of previous-token attention early in training and lightly regularizes predictions toward the closed-form posterior throughout training. Inference uses the two-block model without evaluating auxiliary losses. Across 18 task-pool/seed settings, SAIP reduces mean KL divergence to the bigram posterior by **89.3%**, with at least a 69.8% reduction in every setting. It also strengthens previous-token attention and moves the empirical generalization boundary to smaller task pools. Ablations attribute most of the accuracy gain to posterior regularization, while attention measurements reveal a sharper previous-token circuit. These results show how training can improve the accuracy of Bayesian in-context prediction while changing the attention circuit that implements it.

1 INTRODUCTION

In-context learning allows a transformer to adapt its predictions to a task using examples in its context, without updating its weights (Brown et al., 2020). Understanding this ability requires explaining both the inference strategy a model adopts and how accurately it executes that strategy. Bayesian accounts provide a precise target for the latter question (Xie et al., 2021; Müller et al., 2021; Akyürek et al., 2022): given the evidence in a prompt, how close are the model's predictions to the posterior? A model may favor the appropriate strategy while still leaving much of its attainable accuracy unrealized.

Markov-chain prediction makes this distinction measurable. A sequence reveals an unknown transition matrix, and the Bayes-optimal next-token distribution follows from the observed transition counts. This gives a controlled setting in which to compare learned predictions with an exact statistical target (Bietti et al., 2023; Edelman et al., 2024). In this setting, Gibson et al. (2026) identify four regimes in a two-block transformer, distinguished by memorization versus generalization and unigram versus bigram statistics. Their phase diagram characterizes which Bayes predictor best matches the model as the number of training tasks changes.

That classification leaves an important question open: *how well does the model approximate the predictor that defines its phase?* Being closest to the bigram-generalization posterior imposes no guarantee that the remaining KL divergence is small. Our baseline retains a substantial residual gap even in this regime (Appendix Table 7). This distinction matters for interpreting apparent emergence: a transition to a more suitable inference strategy can occur well before training produces an accurate estimator.

We investigate this gap through the induction-head circuit. Previous-token attention in the first layer supplies the token-pair information needed for the second layer to implement bigram prediction (Olsson et al., 2022; Bietti et al., 2023; Nallamala, 2024). This circuit suggests a testable intervention: guide its formation during training, then measure both posterior accuracy and attention structure in the resulting model. We assess the interventions through both predictions and attention.

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

We introduce *Symmetry-Aware In-context Pretraining* (SAIP; Figure 1) to make this intervention concrete. The design addresses both circuit formation and prediction accuracy. An initial attention bias and a temporary attention loss encourage the first layer to attend to the preceding token. A small posterior-matching loss complements the next-token cross-entropy signal, steering predictions toward the bigram estimator. These interventions act at different points in the computation, which allows us to examine their contributions separately. The attention loss ends after the first 30% of training; the posterior term remains active. SAIP retains the two-block architecture and adds one learned scalar. Access to the closed-form posterior makes this a controlled study of training and mechanism, with a clearly specified statistical target.

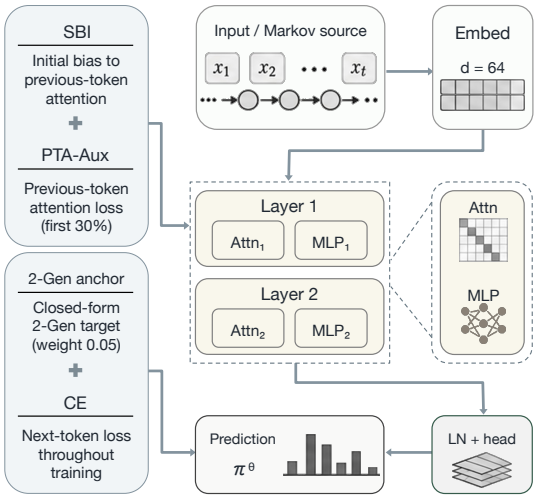


Figure 1: **SAIP overview.** Training interventions in the two-block model. SBI/PTA-Aux guide attention; CE and the 2-Gen anchor supervise predictions.

The experiments establish three findings:

A large part of the residual gap is recoverable through training. Across 18 task-pool/seed settings, SAIP reduces mean KL divergence from 0.0408 to 0.00436, a reduction of 89.3%, with at least 69.8% improvement in every setting (Appendix Table 7). Holding its coefficients fixed while varying vocabulary size and context length retains 74–94% reductions (Table 4). These results quantify improvement within the same model family, beyond identifying its nearest Bayes strategy.

Prediction accuracy and circuit structure respond differently to the intervention. In the ablation setting, posterior regularization alone accounts for 84% of the baseline gap reduction (Figure 3). The full method also raises the measured strength of previous-token attention from 0.40 to 0.80 (Table 1). This division of effects is central to the method's interpretation: the large accuracy gain cannot be attributed to attention guidance alone.

The observed generalization boundary depends on training. SAIP enters the bigram-generalization regime at the smallest task pool we test, while the cross-entropy baseline crosses only at larger pools (Figure 2). It does so with similar measured training time and no additional inference stage. Together with the attention measurements, this shift supports studying phase boundaries as outcomes of a training procedure as well as properties of the task distribution.

The broader question is how training turns a representable inference algorithm into an accurate computation. Our controlled setting exposes a separation between strategy selection, estimator accuracy, and circuit structure. The posterior anchor uses privileged knowledge available in this benchmark, and optimizing toward it directly contributes to the measured gain. Extending the full procedure to settings without that target remains an open problem; the present results identify what each intervention changes where both the statistical target and the mechanism are observable.

2 Related Work

Bayesian inference and strategy selection. Bayesian accounts relate ICL to inference over latent task structure (Xie et al., 2021; Müller et al., 2021). Recent work analyzes Bayesian model averaging and pretraining error (Zhang et al., 2023), learns full parameter posteriors in context (Reuter et al., 2025), and models changing strategy preferences through hierarchical Bayesian inference (Wurgaft et al., 2026). We focus on predictive-posterior accuracy and its mechanistic implementation, rather than parameter-posterior sampling or strategy assignment alone.

Markov learning and algorithmic phases. Markov mixtures expose competition between retrieval and inference strategies (Park et al., 2025); differential subcircuit learning rates can explain memorization–generalization transitions (Nguyen & Nallamala, 2025). We adopt the four-predictor framework of Gibson et al. (2026). Selective-lag and variable-order models broaden the task struc-

ture (D'Angelo et al., 2025; Zhou et al., 2024), while Wei et al. (2026) study Bayesian averaging over uncertain causal dependencies. We retain first-order chains to quantify residual error relative to an explicit posterior.

Training dynamics and circuit formation. Dynamics analyses characterize how ICL emerges in linear attention (Zhang et al., 2025), multi-head softmax regression (He et al., 2025), and Gaussian-mixture classification (Shen et al., 2024). Musat et al. (2025) analyze induction-head formation in a simplified setting. SAIP evaluates an intervention on attention formation and posterior accuracy in the Markov setting.

Mechanisms and targeted training. Induction heads and function-vector heads contribute differently across language-model tasks (Yin & Steinhardt, 2025), while task-vector encoding and decoding can emerge together during training (Han et al., 2024). Mechanistic interventions include attention-score supervision for demonstrated labels (Cho et al., 2025) and weight forgetting to modulate structural ICL (Anand et al., 2025). SAIP targets previous-token attention and combines cross-entropy with posterior regularization. Both SAIP and distillation require the closed-form target.

3 In-Context Markov Learning and Its Four Bayes Phases

Task. Fix a vocabulary $\mathcal{X} = \{1, \ldots, V\}$ and a context length T. A *task* is a row-stochastic transition matrix $P \in \Delta_V^V$ drawn independently for each task identity $k \in \{1, \ldots, K\}$ from a symmetric Dirichlet prior with parameter α. A training sequence is generated by sampling k uniformly and then $x_{1:T} \sim \mathrm{Markov}(P_k)$. The transformer is trained to predict x_{t+1} given $x_{1:t}$ with no access to the task index.

Four Bayes predictors. Following Gibson et al. (2026), four optimal predictors arise depending on whether the model uses (i) 1-point (unigram) or 2-point (bigram) statistics and (ii) whether it memorizes the K tasks or generalizes by Dirichlet smoothing (each predicts the symbol v at position $t{+}1$):

$$\hat{\pi}^{\text{1-Gen}}(v) = \frac{c_t(v) + \alpha}{t + V\alpha},$$

$$\hat{\pi}^{\text{2-Gen}}(v) = \frac{c_t(x_t, v) + \alpha}{c_t(x_t, \cdot) + V\alpha},$$

$$\hat{\pi}^{\text{1-Mem}}(v) = \sum_{k=1}^{K} \Pr(k \mid x_{1:t};\, \text{unigram})\, \pi_k(v),$$

$$\hat{\pi}^{\text{2-Mem}}(v) = \sum_{k=1}^{K} \Pr(k \mid x_{1:t};\, \text{bigram})\, P_k(x_t, v),$$

where $c_t(\cdot)$ and $c_t(\cdot, \cdot)$ are running unigram/bigram counts and π_k is the stationary distribution of P_k. We report the KL divergence to each of the four predictors at the final position of length-T context.

Phase transitions. Following the phase description of Gibson et al. (2026), we distinguish changes in the nearest Bayes predictor as the task-pool size K increases. We use K_1^* for the empirical transition to bigram generalization; the reported reference value is approximately 94. Our discrete sweep estimates this boundary only at the tested pool sizes. Throughout the paper, $\mathrm{div}_{\text{2-Gen}} = \mathrm{KL}(\hat{\pi}^{\text{2-Gen}} \,\|\, \pi^{\theta})$ measures discrepancy from the bigram-generalization posterior, including settings where another predictor is closer. Its mathematical minimum is zero. A positive residual divergence therefore indicates approximation error, even when 2-Gen is the nearest of the four predictors.

4 Method: Symmetry-Aware In-Context Pre-Training

The model follows the two-block transformer configuration of Gibson et al. (2026): token embeddings of width $d{=}64$, causal single-head self-attention with rotary positional embeddings (Su et al., 2024), a GELU feed-forward sublayer of width $4d$, LayerNorm (Ba et al., 2016), and a tied output head. SAIP retains this depth and width, adds one scalar attention-bias parameter, and modifies the training objective. The auxiliary losses are not evaluated during inference.

Symmetry-Breaking Initialization (SBI). We introduce a learnable scalar $\alpha_{\mathrm{SBI}} \in \mathbb{R}$ initialized to 0.5. Before the layer-one attention softmax, we add $\alpha_{\mathrm{SBI}}\mathbf{M}_{\mathrm{prev}}$ to the logits, where $\mathbf{M}_{\mathrm{prev}} \in \{0,1\}^{T\times T}$ has ones on the first lower sub-diagonal. This bias favors the immediately preceding token at initialization. The scalar is optimized jointly with the other parameters and is not constrained to remain positive or to decay to zero. SBI thus specifies an initial preference for a circuit motif rather than a fixed attention pattern.

Posterior-Targeted Attention auxiliary loss (PTA-Aux). At training step $t<t_{\mathrm{decay}}{=}0.3T_{\mathrm{train}}$ we add a KL term between the realized layer-one attention map $\mathbf{A}_1 \in \Delta_T^T$ and the ideal previous-token pattern $\mathbf{M}_{\mathrm{prev}}$:

$$\mathcal{L}_{\mathrm{PTA}}(t) = \tfrac{1}{2}\lambda_{\mathrm{PTA}}\left(1+\cos\tfrac{\pi t}{t_{\mathrm{decay}}}\right)\mathrm{KL}(\mathbf{M}_{\mathrm{prev}} \,\|\, \mathbf{A}_1)\,,$$

with $\lambda_{\mathrm{PTA}}{=}0.05$. This term encourages the previous-token pattern during early training. By construction, $\mathcal{L}_{\mathrm{PTA}}(t) = 0$ for $t \geq t_{\mathrm{decay}}$, so the attention-specific regularizer is absent thereafter. PTA-Aux is the component that directly regularizes an internal attention map.

2-Gen anchor. We add a small constant KL term between the model's output distribution and the closed-form 2-Gen Bayes predictor $\hat{\pi}^{\text{2-Gen}}$, computed in-batch from the same sequences:

$$\mathcal{L}_{\mathrm{2G}} = \lambda_{\mathrm{2G}}\,\mathrm{KL}\big(\hat{\pi}^{\text{2-Gen}} \,\|\, \pi^{\theta}\big)\,,$$

with $\lambda_{\mathrm{2G}}{=}0.05$, selected from the sensitivity sweep (Appendix D). The anchor directly optimizes the divergence used for evaluation, while cross-entropy continues to supervise observed next tokens. Consequently, reductions in this metric reflect posterior supervision as well as any benefit from attention guidance. The ablations evaluate these contributions separately.

Final objective. Combining the three components,

$$\mathcal{L}(t) = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{PTA}}(t) + \mathcal{L}_{\mathrm{2G}}.$$

For $t \geq t_{\mathrm{decay}}$, the objective becomes $\mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{2G}}$. The anchor coefficient remains constant, so this objective differs from cross-entropy alone. The training objective explicitly combines empirical next-token prediction with posterior supervision. The experiments assess its predictive effects alongside the attention changes induced during the early regularization interval.

Hyperparameter robustness. The coefficients $\alpha_{\mathrm{SBI},0}{=}0.5$, $\lambda_{\mathrm{PTA}}{=}0.05$, and $\lambda_{\mathrm{2G}}{=}0.05$ are fixed across the main K sweep. Setting both auxiliary coefficients to zero and fixing $\alpha_{\mathrm{SBI}} \equiv 0$ recovers the cross-entropy baseline. The anchor sweep exhibits decreasing divergence with diminishing improvements beyond 0.05; the remaining sensitivity analyses are reported in Appendix D.

5 Experiments

Setup. We use the two-block Markov-ICL benchmark of Gibson et al. (2026): an Attn–MLP–Attn–MLP transformer ($d{=}64$, single head, RoPE) predicts the next token in length-$T{=}128$ sequences sampled from pools of symmetric-Dirichlet($\alpha{=}1$) Markov chains over $V{=}10$ symbols. The main sweep comprises nine pool sizes and two seeds, giving 18 (K, seed) settings. SAIP uses fixed coefficients $\alpha_{\mathrm{SBI},0}{=}0.5$ and $\lambda_{\mathrm{PTA}}{=}\lambda_{\mathrm{2G}}{=}0.05$. We evaluate the four posterior divergences on a held-out batch of 512 sequences. Additional seed repetitions and configuration details are specified in the relevant tables and Appendix A. Training times are measured on an RTX 4090.

Baselines. Table 1 compares SAIP with seven methods and two reference predictors. The attention-based methods are the Two-Block baseline (Edelman et al., 2024; Gibson et al., 2026), a Single-Block depth ablation (Bietti et al., 2023), Posterior Distillation (Hahn & Goyal, 2023), an Induction-Head Primer retaining the initial attention bias (Olsson et al., 2022; Nallamala, 2024), and a Mechanistic Curriculum varying K during training (Nallamala, 2024). We additionally evaluate a compact Mamba implementation (Bondaschi et al., 2026) and order-1 CTW with KT smoothing (Zhou et al., 2024). Random initialization and the exact 2-Gen posterior provide reference predictions. Mamba is an implementation-specific comparison, with limitations described in Appendix J. All reported outcomes refer to our implementations and evaluation protocol.

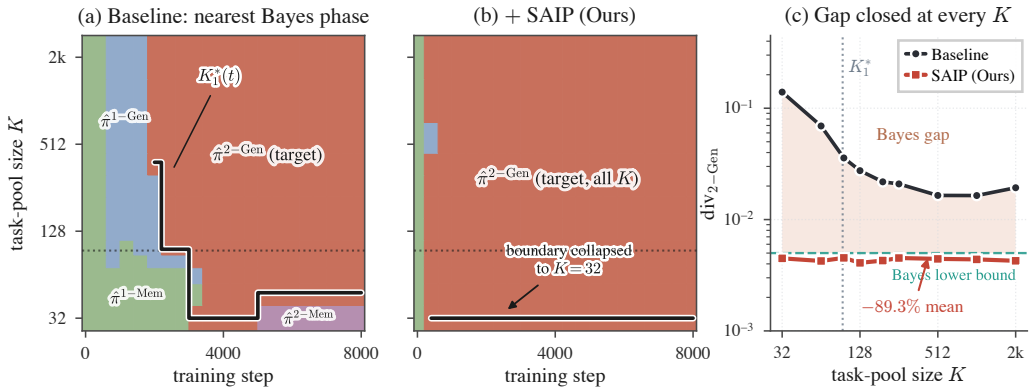


Figure 2: **Strategy assignment and posterior divergence during training. (a,b)** Nearest Bayes predictor over (K, step) for the cross-entropy baseline and SAIP. The black curve marks the empirical transition to 2-Gen; the dotted reference is $K_1^* \approx 94$ from Gibson et al. (2026). SAIP reaches the 2-Gen regime at smaller tested pools. **(c)** Divergence to the bigram posterior across K. The main 18-setting evaluation gives an 89.3% reduction in mean divergence (Appendix Table 7).

Main results. Figure 2 reports both strategy assignment and prediction error. Panels (a,b) identify the nearest Bayes predictor over task-pool size and training step. SAIP reaches the 2-Gen regime at smaller tested pools than the cross-entropy baseline. Panel (c) shows lower divergence to the bigram posterior across the sweep. These observations distinguish two effects: a change in strategy assignment at smaller pools, and improved approximation where the baseline already favors 2-Gen. At $K{=}128$, the largest change among the four posterior divergences is to 2-Gen; the remaining divergences vary comparatively little (Table 3).

Table 1 reports prediction error, phase assignment, attention statistics, and whether cross-entropy is retained. SAIP and Posterior Distillation have the lowest divergences among the trained methods in this comparison. Both optimize toward the posterior used for evaluation: Posterior Distillation assigns that loss unit weight, whereas SAIP uses coefficient 0.05 alongside cross-entropy and attention guidance. Their difference therefore concerns the training signal and the resulting solution, rather than access to the Bayesian target. Initialization alone and the depth-ablated model yield higher residual divergence. The full per-K results are provided in Appendix Table 7.

Per-setting improvements and convergence. SAIP improves divergence in all 18 evaluated (K, seed) settings, with reported reductions of 70–97% and an aggregate mean-divergence reduction of 89.3% (Appendix Table 7). The trajectories in Figure 4b show faster reduction and a lower final divergence for SAIP within the training budget. End-of-training attention maps show greater concentration on the previous-token diagonal (Appendix Figure 9), consistent with the circuit interpretation of Olsson et al. (2022); Nallamala (2024).

Sensitivity to aggregation and task-pool range. Table 2 reanalyzes the paired 18-setting records. The ratio-of-means reduction weights settings by their baseline divergence, whereas the mean paired reduction gives each setting equal weight. The latter is 82.7%, compared with the main 89.3% aggregate. Restricting the comparison to $K \geq 512$ still yields a 74.9% ratio-of-means reduction. Thus the large-error small-pool settings contribute to the aggregate magnitude, but the improvement persists when they are excluded. These summaries use the same trained models and quantify sensitivity to evaluation aggregation.

In-context behavior across positions. The primary metric evaluates the final context position. To assess whether the improvement extends to earlier positions, we also measure $\text{div}_{2\text{-Gen}}$ throughout the context (Appendix Figure 18). SAIP has lower divergence across the evaluated positions, indicating that its advantage is not restricted to the final-token read-out.

Table 1: **Comparison of methods and reference predictors.** Mean $\text{div}_{\text{2-Gen}}$ uses $K \in \{32, 64, 128, 2048\}$ (5 seeds for SAIP/Posterior Distillation, 2 for the other trained methods). K_1^* is the smallest tested pool classified as 2-Gen; $\varphi_\delta^{(1)}$ measures previous-token attention; n_A is the effective attended-position count; CE indicates retention of cross-entropy. The full 18-setting baseline/SAIP sweep is reported separately in Appendix Table 7.

Method	$\text{div}_{\text{2-Gen}}$	$K_1^*\downarrow$	$\varphi_\delta^{(1)}\uparrow$	$n_A\downarrow$	CE
$\hat{\pi}^{\text{2-Gen}}$ Bayes oracle	0.000	–	–	–	–
Random init	0.649	>2k	0.04	–	–
Single-Block Tr. '23[a]	0.158	>2k	0.04	128	✓
Two-Block Tr. '24[b]	0.064	64	0.40	22.6	✓
Mech. Curriculum '24[c]	0.061	64	0.40	15.1	✓
Ind.-Head Primer '22[d]	0.057	64	0.40	44.4	✓
Mamba 1-layer SSM '26[e]	0.134	–	–	–	✓
CTW (order-1, KT) '26[f]	0.020	≤32	–	–	–
Post. Distillation '23[g]	0.007	≤32	0.59	7.1	×
SAIP (Ours)	**0.005**	**≤32**	**0.80**	**5.5**	✓

[a] Bietti et al. (2023); Makkuva et al. (2025); Ekbote et al. (2026). [b] Edelman et al. (2024); Gibson et al. (2026); Lepage et al. (2025); D'Angelo et al. (2025). [c] Nallamala (2024). [d] Olsson et al. (2022); Nallamala (2024). [e] Bondaschi et al. (2026): compact pure-PyTorch SSM, both 1/2-layer plateau at ≈ 0.13 (App. J); cross-paradigm reference. [f] Zhou et al. (2024), KT(α=0.5). [g] Hahn & Goyal (2023): has greater observed seed variability ($\approx 9\times$ the SAIP seed s.d., with one high-error run among five; App. G). **Bold** = best non-oracle. SAIP uses anchor λ_{2G}=0.05.

Table 2: **Paired improvement under alternative aggregations.** The paired λ_{2G}=0.05 sweep. Entries are percentage divergence reductions. For paired divergences (b_i, s_i), pooled reduction is $100(1 - \sum_i s_i / \sum_i b_i)$; equal-weight reduction averages $100(1 - s_i/b_i)$. The nested subsets are descriptive sensitivity checks.

Task pools	Pairs	Pooled	Equal-weight	Minimum
$K \geq 32$	18	89.3	82.7	69.8
$K \geq 96$	14	80.7	79.1	69.8
$K \geq 512$	6	74.9	74.1	69.8

Ablations. We isolate each SAIP component by training six leave-one-out / leave-one-in variants (mean $\text{div}_{\text{2-Gen}}$ across $K \in \{32, 128, 512, 2048\}$, 2 seeds), summarized in Figure 3.

Figure 3 identifies the posterior anchor as the largest contributor to divergence reduction in the ablation setting. At λ_{2G}=0.02, anchor-only training reduces mean divergence from 0.051 to 0.008, approximately 84%. Full SAIP yields 0.007, while removing the anchor gives 0.059. The remaining components therefore provide a smaller additional gain on this metric. Increasing λ_{2G} over $[0.005, 0.2]$ gives divergences $0.017/0.012/0.007/\mathbf{0.005}/0.005/0.004$, with diminishing improvement beyond the main setting 0.05. These comparisons characterize component effects at 0.02, distinct from the main 0.05 setting. They show that posterior supervision accounts for most of the predictive improvement, while the attention measurements evaluate the accompanying circuit changes.

Mechanistic verification: K_1^* shift and order parameters. Table 1 and Figure 4 report complementary attention measurements. SAIP is classified as 2-Gen at the smallest tested pool, giving $K_1^* \leq 32$, whereas the discrete baseline estimate is 64. These estimates should be distinguished from the approximately 94 reference boundary reported by Gibson et al. (2026). At K=128, SAIP has previous-token attention $\varphi_\delta^{(1)}$=0.80, compared with 0.59 for Posterior Distillation and 0.40 for the baseline and Induction-Head Primer. Its effective attended-position count is also lower (n_A=5.5, compared with 7.1, 22.6, and 44.4). The Primer comparison suggests that the initial bias alone does not reproduce the attention concentration of full SAIP. In Table 3, the 2-Gen divergence decreases from 0.031 to 0.005, while the other divergences remain approximately 0.18. The attention pattern persists after PTA-Aux ends (Appendix Figure 17). Together these measurements associate improved posterior prediction with earlier formation and greater concentration of previous-token attention.

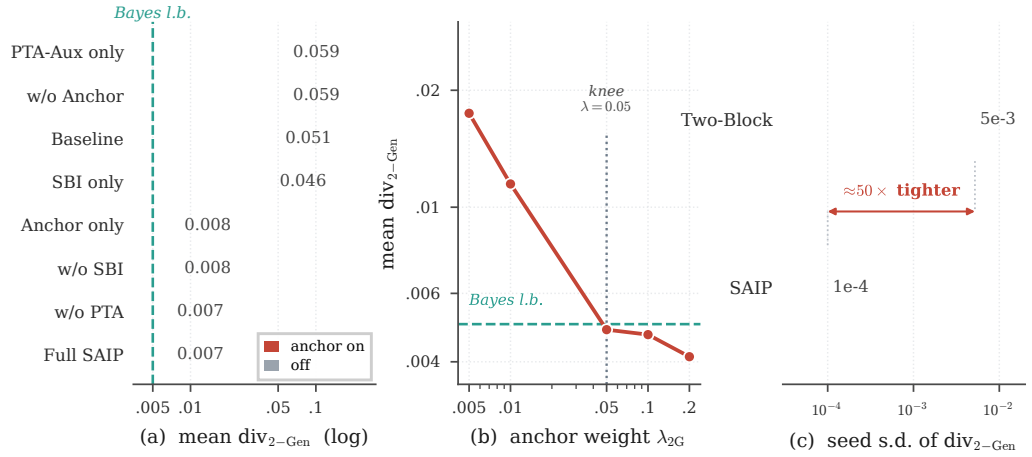


Figure 3: **Component ablation and sensitivity.** **(a)** Mean $\text{div}_{\text{2-Gen}}$ for component variants (log scale; red denotes an active anchor, grey its absence). Anchor-containing variants have lower divergence. **(b)** Divergence decreases with anchor weight, with diminishing improvements beyond the main coefficient 0.05. **(c)** Seed standard deviation for the Two-Block baseline and SAIP.

Table 3: **Targeted intervention.** $\text{KL}(\hat{\pi}^{\bullet} \,\|\, \pi^{\theta})$ to each Bayes predictor at $K{=}128$ (mean over seeds). The largest change is in the $\hat{\pi}^{\text{2-Gen}}$ column; other divergences remain approximately 0.18.

Method	$\hat{\pi}^{\text{1-Gen}}$	$\hat{\pi}^{\text{2-Gen}}$	$\hat{\pi}^{\text{1-Mem}}$	$\hat{\pi}^{\text{2-Mem}}$
Two-Block Tr.	0.176	0.031	0.175	0.178
IH Primer	0.184	0.027	0.182	0.172
Post. Distill.	0.175	0.008	0.177	0.187
SAIP (Ours)	0.180	**0.005**	0.181	0.181

Table 4: **Generalization** ($K{=}128$; mean $\text{div}_{\text{2-Gen}}$). SAIP closes 74–94% of the baseline gap across V and T without re-tuning.

	setting	Two-Block	SAIP
vocab. V	$V{=}4$	0.0103	**0.0022**
	$V{=}10$	0.0278	**0.0046**
	$V{=}16$	0.1144	**0.0068**
ctx. T	$T{=}64$	0.0288	**0.0075**
	$T{=}128$	0.0278	**0.0046**
	$T{=}256$	0.0256	**0.0045**

Representation and read-out evidence. Figure 5 connects the attention measurements to three complementary diagnostic probes. Previous-token identity explains 0.437 of layer-one residual variance for SAIP, compared with 0.235 for the baseline and 0.374 for Posterior Distillation. Yet the predictions decoded after layer one have similar posterior divergence across methods. The output difference becomes pronounced after layer two, where baseline divergence is 0.0265, compared with approximately 0.0044–0.0045 for the posterior-supervised methods. Stronger organization of the intermediate representation therefore accompanies improved final prediction, without implying that an accurate posterior can already be decoded from layer one. Panel (c) additionally reports posterior divergence along straight parameter interpolations. The Posterior Distillation–SAIP path rises above both low-error endpoints, showing that similar output accuracy does not imply low divergence along their direct parameter interpolation. These diagnostic runs are separate from the main 18-setting sweep.

Depth ablation. The single-block model has higher divergence in this experiment, increasing from 0.148 to 0.172 as K grows. This is consistent with the role of depth in the circuit analyses of Gibson et al. (2026); Edelman et al. (2024), within the tested width and training budget. SAIP improves the two-block model without adding layers or attention heads.

Statistical robustness and generalization. With coefficients fixed, we vary vocabulary size V and context length T at $K{=}128$ (Table 4). Relative divergence reductions range from 74 to 94%, including 94% at $V{=}16$. This establishes transfer across the tested configurations without per-setting retuning. Across five seeds at the default setting, the reported standard deviations are 1.0×10^{-4} for SAIP and 5.2×10^{-3} for the baseline, approximately a $50\times$ difference. These finite-sample

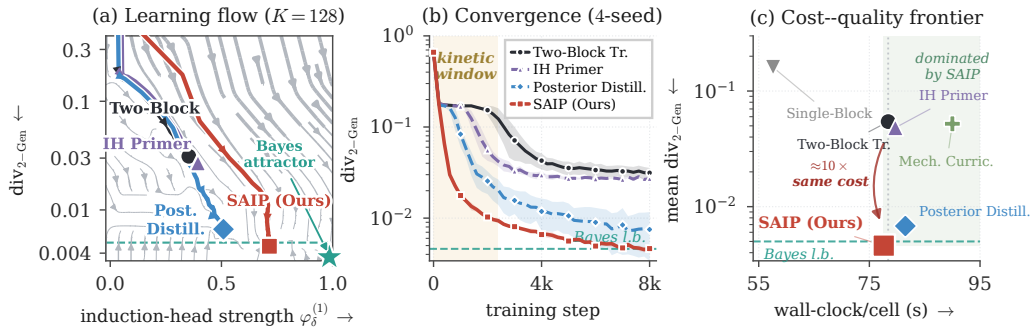


Figure 4: **Training trajectories, attention, and cost** ($K=128$). **(a)** Trajectories in the $(\varphi_\delta^{(1)}, \mathrm{div}_{\text{2-Gen}})$ plane. Grey streamlines summarize pooled checkpoint displacements; bold curves show method means (open markers: initialization; filled markers: final values). SAIP ends with higher previous-token attention than the other displayed methods. **(b)** Divergence over training (4-seed mean±s.d.). Shading marks the interval in which PTA-Aux is active. **(c)** Measured training time and final divergence. SAIP combines lower divergence with similar median time to the Two-Block baseline; these timings are observations on shared hardware.

estimates describe variability in the measured runs. Appendix Figures 12, 11, and 18 provide the seed distributions, attention evolution, and position-wise divergence.

Robustness to hyperparameters. Within the tested ranges, divergence is most sensitive to the posterior-anchor weight (Appendix Figure 7). Increasing λ_{2G} from 0.005 reduces divergence from 1.7×10^{-2}, with diminishing gains beyond 0.05. The PTA-Aux weight, initial bias, and decay fraction produce smaller changes near the chosen configuration. The joint 5×5 loss-weight sweep similarly shows a broad region of low divergence at larger anchor weights (Appendix Figure 8). The anchor weight is therefore the main source of predictive sensitivity in these sweeps, while the remaining coefficients specify the early attention intervention.

Cost. The measured median training times are $77.6\,\mathrm{s}$ for SAIP and $78.4\,\mathrm{s}$ for the Two-Block baseline (Appendix Table 9). The anchor reuses model outputs, while PTA-Aux requires an attention computation during the first 30% of training. The observed timing difference is small on the shared GPU, and the timing protocol is described in Appendix G. Figure 4c reports the measured time–divergence trade-off.

6 Discussion

Training dependence of prediction and mechanism. The results show that changing the training procedure substantially reduces posterior divergence within the same two-block model family. This supports an optimization contribution to the residual gap. The attention trajectories additionally associate the improvement with earlier formation and greater concentration of previous-token attention. However, SAIP also changes the objective through persistent posterior supervision, so the experiment does not isolate a purely kinetic effect under a common objective. The shift in the empirical phase boundary should be interpreted with this distinction in mind. The interpolation and logit-lens probes provide further descriptions of the learned solutions, rather than independent proofs of the proposed causal mechanism.

Comparison to Posterior Distillation. Posterior Distillation and SAIP obtain low posterior divergence with different attention statistics. Across five seeds, distillation has approximately $9\times$ greater standard deviation and one high-error run; its previous-token attention is 0.59, versus 0.80 for SAIP. Both require posterior access. Representational differences also arise between ICL and supervised fine-tuning (Confavreux et al., 2026), a distinct adaptation comparison. Our measurements similarly motivate separating output accuracy from internal computation; they do not establish that attention differences cause the stability difference.

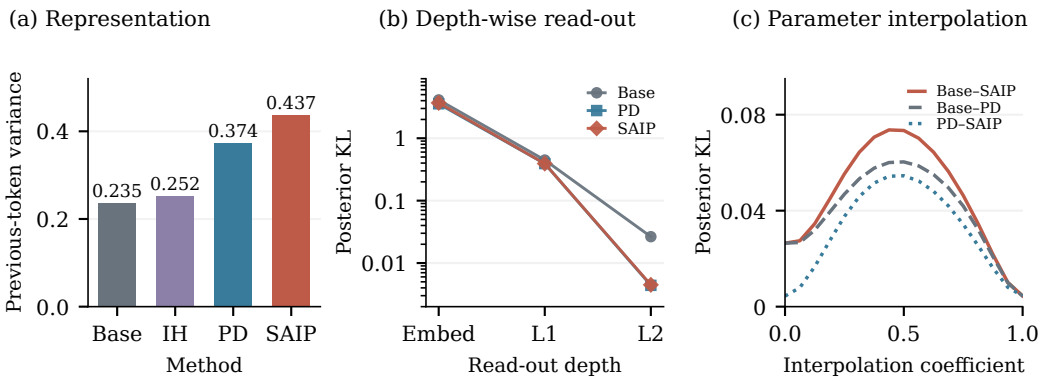


Figure 5: **Representation, prediction, and parameter-space diagnostics** (K=128). **(a)** Between-class residual variance divided by total variance for previous-token identity (Base: Two-Block; IH: Induction-Head Primer; PD: Posterior Distillation). **(b)** Posterior KL decoded at three depths with the final LayerNorm and tied head. **(c)** Posterior KL along straight parameter interpolations between model pairs, from the first model at coefficient zero to the second at one. The barriers describe the tested paths, not topological separation of solutions.

Posterior supervision and solution selection. The anchor supplies a structured posterior target derived from the training sequence. It uses no additional observations, but provides supervision that next-token cross-entropy leaves implicit. Its effect is substantial: anchor-only training accounts for approximately 84% of the ablation reduction. The regularizer therefore favors predictions closer to the posterior. Whether it also selects among distinct optimization basins remains unresolved: the interpolation probes establish barriers along straight paths, while more general connectivity is unmeasured. Distinguishing the effects of the additional target from those of early attention guidance requires matched component comparisons.

Joint interpretation of the additional evidence. The two analyses address complementary explanations of the gain. Table 2 shows that improvement persists beyond the smallest pools and under equal setting weights. Figure 5 shows different intermediate representations despite similar final posterior accuracy. Together, these results support evaluating the full procedure through both paired predictive improvements and the internal computation learned under posterior and attention supervision.

Limitations and scope. The experiments establish performance in a first-order Markov benchmark with a known posterior. Extending SAIP to settings without that posterior requires a different target and new evidence of effectiveness. An expert or moving-average teacher is a possible direction, not a validated substitute. The study also uses one model width (d=64) and two-block depth; the depth ablation does not establish robustness across scales. The previous-token prior is appropriate to this task structure, and its suitability for other dependencies must be tested. Finally, the main sweep uses two seeds, while several diagnostic probes use a single seed. Larger matched comparisons, including anchor-only training at the main 0.05 coefficient, would strengthen attribution of the observed gains. A complementary test would remove the anchor after matched training checkpoints and track both divergence and attention. Persistence would support a lasting change in the learned computation; deterioration would indicate dependence on continued posterior supervision.

7 Conclusion

SAIP improves Bayesian in-context prediction in two-block Markov transformers through early attention guidance and persistent posterior regularization. It reduces mean KL divergence by 89.3%, with improvements across all paired settings and under alternative aggregations. The posterior anchor accounts for most of the accuracy gain, while attention and representation probes characterize changes in the learned computation. These findings distinguish strategy assignment, prediction accuracy, and internal representation as complementary dimensions of in-context learning. The closed-form target limits the demonstrated scope to this controlled benchmark. Extending the approach without posterior access and isolating the contribution of circuit guidance at larger scales remain open questions.

AI USE STATEMENT

OpenAI Codex assisted with manuscript revision, including the interpretation and presentation of existing results, feedback on experimental comparisons, and suggestions for follow-up analyses. It also assisted with drafting and editing text, literature searches, bibliographic checks, reference formatting, and figure design and placement. AI-assisted Python scripts recomputed descriptive statistics from existing experimental records and generated additional visualizations, including aggregation-sensitivity summaries and representation, read-out, and parameter-interpolation plots. Checks performed in this revision workflow included matching paired records and configurations, comparing numerical summaries with source files, consulting original publication pages for bibliographic metadata, and inspecting compiled figures and document layout. The authors are responsible for the accuracy, originality, and final content of the submission, including all AI-assisted text, interpretations, code, and figures.

ETHICS STATEMENT

The experiments reported here use procedurally sampled Markov sequences in a controlled prediction benchmark and do not involve human participants or personal data. The work studies how training objectives affect posterior approximation and internal representations; it does not evaluate a deployed decision system or establish fairness or safety properties in real-world applications. The posterior anchor assumes access to a statistical target that is available in this benchmark. Applying the method to consequential decisions would require separate validation of the data assumptions, target quality, and relevant harms. For research integrity, we distinguish measured improvements from causal interpretations, identify implementation-specific baseline limitations, and report the scope of the diagnostic probes. AI assistance in manuscript preparation and analysis is described above.

REPRODUCIBILITY STATEMENT

Section 3 specifies the task distribution and posterior predictors, and Section 4 defines the training interventions and loss schedules. Appendix A records the model, optimizer, compute, and seed settings; Appendix B distinguishes method configurations and the main and ablation coefficients. Appendix C provides the paired task-pool/seed results, with sensitivity and attention definitions in Appendices D and E. Training pseudocode and implementation details appear in Appendix L, and Appendix M documents evaluation consistency and remaining nondeterminism. Computational-cost measurements and numerical diagnostic values are provided in Appendices G and N. The aggregation formulas in Table 2 specify how the alternative improvement summaries are computed.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Suraj Anand, Michael Lepori, Jack Merullo, and Ellie Pavlick. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. In *International Conference on Learning Representations*, volume 2025, pp. 50280–50306, 2025.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Anooshka Bajaj, Deven Mahesh Mistry, Sahaj Singh Maini, Yash Aggarwal, Billy Dickson, and Zoran Tiganj. Temporal dependencies in in-context learning: The role of induction heads. *arXiv preprint arXiv:2604.01094*, 2026.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Makkuva. From markov to laplace: How mamba in-context learns markov

chains. In *International Conference on Learning Representations*, volume 2026, pp. 89166–89196, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hakaze Cho, Peng Luo, Mariko Kato, Rin Kaenbyou, and Naoya Inoue. Mechanistic fine-tuning for in-context learning. In *Proceedings of the 8th Blackboxnlp Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 330–357, 2025.

Basile Confavreux, Aaditya Singh, Jin Hwa Lee, Amaury Sabran, and Andrew Saxe. Comparing the learning dynamics of in-context learning and fine-tuning in language models. In *International Conference on Learning Representations*, volume 2026, pp. 157425–157437, 2026.

Francesco D'Angelo, Nicolas Flammarion, et al. Selective induction heads: How transformers select causal structures in context. In *International conference on learning representations*, volume 2025, pp. 86752–86790, 2025.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. *Advances in neural information processing systems*, 37:64273–64311, 2024.

Chanakya Ekbote, Ashok Vardhan Makkuva, Marco Bondaschi, Nived Rajaraman, Michael Gastpar, Jason Lee, and Paul Liang. What one cannot, two can: Two-layer transformers provably represent induction heads on any-order markov chains. *Advances in Neural Information Processing Systems*, 38:37523–37603, 2026.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International conference on machine learning*, pp. 3259–3269. PMLR, 2020.

Cole Gibson, Wenping Cui, and Gautam Reddy. Distinct mechanisms underlying in-context learning in transformers. *arXiv preprint arXiv:2604.12151*, 2026.

Michael Hahn and Navin Goyal. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*, 2023.

Seungwook Han, Jinyeop Song, Jeff Gore, and Pulkit Agrawal. Emergence and effectiveness of task vectors in in-context learning: An encoder decoder perspective. *arXiv preprint arXiv:2412.12276*, 2024.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. *arXiv preprint arXiv:2503.12734*, 2025.

Alexander Ku, Thomas L Griffiths, and Stephanie Chan. An evolutionary perspective on modes of learning in transformers. In *International Conference on Learning Representations*, volume 2026, pp. 80997–81009, 2026.

Simon Lepage, Jeremie Mary, and David Picard. Markov chain estimation with in-context learning. *arXiv preprint arXiv:2508.03934*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Nischal Mainali and Lucas Teixeira. Exact learning dynamics of in-context learning in linear transformers and its application. *Proceedings of Machine Learning Research*, 282:333–368, 2025.

Ashok Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A curious case of single-layer transformers. In *International Conference on Learning Representations*, volume 2025, pp. 853–879, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Tiberiu Musat, Tiago Pimentel, Lorenzo Noci, Alessandro Stolfo, Mrinmaya Sachan, and Thomas Hofmann. On the emergence of induction heads for in-context learning. *arXiv preprint arXiv:2511.01033*, 2025.

Gautam Reddy Nallamala. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *International conference on learning representations*, volume 2024, pp. 19226–19239, 2024.

Alex Nguyen and Gautam Reddy Nallamala. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 8677–8697, 2025.

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.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 66381–66433, 2025.

Arik Reuter, Tim GJ Rudner, Vincent Fortuin, and David Rügamer. Can transformers learn full bayesian inference in context? *arXiv preprint arXiv:2501.16825*, 2025.

Wei Shen, Ruida Zhou, Jing Yang, and Cong Shen. On the training convergence of transformers for in-context classification of gaussian mixtures. *arXiv preprint arXiv:2410.11778*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jianzhe Wei, Siyu Chen, Jianliang He, and Zhuoran Yang. How transformers learn causal structures in-context: Explainable mechanism meets theoretical guarantee. In *International Conference on Learning Representations*, volume 2026, pp. 91135–91176, 2026.

Daniel Wurgaft, Ekdeep S Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah Goodman. In-context learning strategies emerge rationally. *Advances in Neural Information Processing Systems*, 38:118425–118485, 2026.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Cong Xu, Guoliang Li, Jun Wang, and Wei Zhang. Markovian pre-trained transformer for next-item recommendation. *arXiv preprint arXiv:2601.08275*, 2026.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? *arXiv preprint arXiv:2502.14010*, 2025.

Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*, 2023.

Ruida Zhou, Chao Tian, and Suhas Diggavi. An information-theoretic approach to understanding transformers' in-context learning of variable-order markov chains. *arXiv preprint arXiv:2410.05493*, 2024.

SUPPLEMENTARY MATERIAL (APPENDIX)

This appendix provides experimental configurations, additional evaluations, and diagnostic probes. Sections A and B specify the training setup and method variants. Section C reports the full baseline/SAIP sweep. Sections D–F examine hyperparameter sensitivity, attention statistics, and training trajectories. Sections G and H cover computational cost and context-dependent behavior. Sections I and N describe interpolation and logit-lens measurements, and Section K discusses the relationship to other settings.

A EXACT TRAINING CONFIGURATION

We use the Markov-chain ICL configuration of Gibson et al. (2026); Table 5 summarizes the task, model, and optimization settings.

Table 5: Task, model, and optimization settings for the main evaluation.

Task		**Model**	
Vocabulary size V	10	Hidden dim d	64
Context length T	128	Heads	1
Dirichlet concentration α	1	Feed-forward width d_{ff}	$4d = 256$
Task pool size K	$\{32, \ldots, 2048\}$ (9 values)	Activation	GELU
Evaluation batch	512	Position encoding	RoPE (Su et al., 2024)
Optimization		**SAIP-specific**	
Optimizer	AdamW (Loshchilov & Hutter, 2017)	$\alpha_{\text{SBI},0}$ (initial)	0.5 (learnable)
Learning rate η	3×10^{-4}	λ_{PTA}	0.05
(β_1, β_2)	$(0.9, 0.95)$	λ_{2G} (main)	0.05
Weight decay	0	t_{decay}	$0.3\,T_{\text{train}}$
Grad-clip max-norm	1.0	PTA-Aux decay shape	$\frac{1}{2}(1+\cos\frac{\pi t}{t_{\text{decay}}})$
Batch size	128	Anchor target	closed-form $\hat{\pi}^{\text{2-Gen}}$
Train steps T_{train}	8000	Anchor direction	$\text{KL}(\hat{\pi}^{\text{2-Gen}} \,\|\, \pi^{\theta})$
Seeds	$\{0, 1\}$; $\{0..4\}$ for SAIP/PD div	Embedding tying	input/output tied head
Compute	$1\times$ RTX 4090	Layer-norm	pre-norm in each block

Relation to the reference configuration. The architecture, task distribution, and optimizer follow the setting of Gibson et al. (2026). SAIP adds the scalar α_{SBI} and the losses $\mathcal{L}_{\text{PTA}}$ and $\mathcal{L}_{\text{2G}}$. Fixing the attention-bias scalar at zero and disabling both losses recovers the cross-entropy baseline. Merely initializing the scalar to zero does not impose that constraint during subsequent optimization.

B METHOD CONFIGURATIONS

Table 6 specifies the initial attention bias, auxiliary coefficients, architecture, and objective for each method. The trainable methods use the shared optimization implementation and step budget, with method-specific changes shown in the table. The curriculum additionally varies the task-pool schedule.

Main operating point. The default ablation configuration uses $\lambda_{\text{2G}}{=}0.02$, while the main evaluation uses 0.05, selected from the sensitivity sweep. Results from these settings should be distinguished when comparing component effects. Table 1 reports the main method comparison; Table 7 reports the separate 18-setting baseline/SAIP sweep.

C FULL TASK-POOL AND SEED EVALUATION GRID

Table 7 lists every (K, seed) cell of the main 18-cell sweep ($K{\in}\{32, 64, 96, 128, 192, 256, 512, 1024, 2048\}$, 2 seeds) for the cross-entropy baseline and SAIP (main, $\lambda_{\text{2G}}{=}0.05$). All values are $\text{KL}(\hat{\pi}^{\text{2-Gen}}\|\pi^{\theta})$ at the final position $t{=}T$, averaged over a held-out evaluation batch of 512 sequences.

Table 6: Training configurations for the compared methods.

Method	$\alpha_{\mathrm{SBI},0}$	λ_{PTA}	λ_{2G}	Arch.	CE	Distill
Two-Block baseline	0	0	0	2-block	✓	
Posterior Distillation	0	0	–	2-block		✓(full-wt)
Induction-Head Primer	0.5	0	0	2-block + PTA bias init	✓	
Single-Block	0	0	0	1-block (depth ablation)	✓	
Mech. Curriculum	0	0	0	2-block, K-curriculum	✓	
SAIP w/ SBI only	0.5	0	0	2-block	✓	
SAIP w/ PTA only	0	0.05	0	2-block	✓	
SAIP w/ Anchor only	0	0	0.02	2-block	✓	
SAIP w/o SBI	0	0.05	0.02	2-block	✓	
SAIP w/o PTA	0.5	0	0.02	2-block	✓	
SAIP w/o Anchor	0.5	0.05	0	2-block	✓	
SAIP (full, default)	0.5	0.05	0.02	2-block	✓	
SAIP (Ours, main)	**0.5**	**0.05**	**0.05**	**2-block**	✓	

Table 7: Full 18-cell grid. Values $\times 10^3$. Per-cell reduction (Ours vs. baseline) shown in the last two columns.

	Two-Block Baseline		SAIP (Ours)		% reduction	
K	seed 0	seed 1	seed 0	seed 1	seed 0	seed 1
32	141	138	4	5	97	97
64	59	79	4	4	93	95
96	34	37	5	5	87	88
128	27	28	4	4	85	85
192	22	21	4	4	81	80
256	19	23	4	5	76	80
512	17	16	4	4	74	73
1024	15	18	4	5	71	75
2048	14	24	4	4	70	83
mean	40.8		4.36		**89.3**	

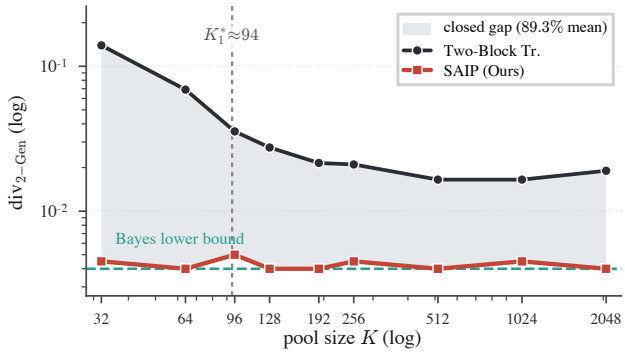


Figure 6: **Divergence reduction across task-pool size** (2-seed means; Appendix Table 7). SAIP has lower divergence than the cross-entropy baseline at each tested K. Shading indicates the difference between methods. The reduction in mean divergence is 89.3%, with 97% at $K{=}32$.

Interpretation. Figure 6 summarizes the per-pool divergences. Reductions tend to be larger at smaller K, where the baseline discrepancy is greatest. SAIP remains near 4×10^{-3} across the sweep, with the largest reported reduction (97%) at $K{=}32$.

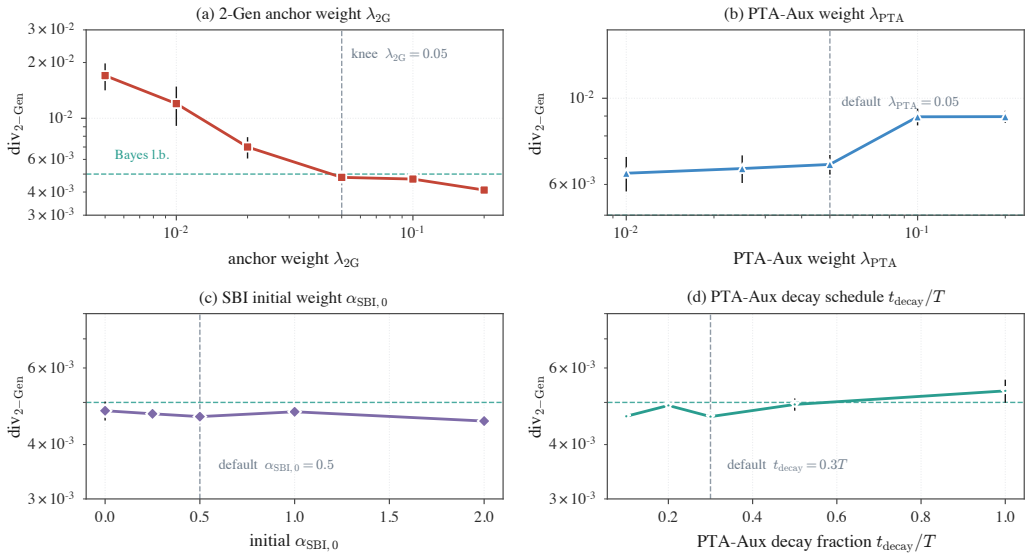


Figure 7: **Hyperparameter sensitivity.** Each panel varies one coefficient while holding the others fixed (K=128, V=10, T=128, 2 seeds). **(a)** Anchor weight: divergence decreases, with diminishing gains beyond 0.05. **(b)** PTA-Aux weight: low divergence over $[0.01, 0.05]$, followed by degradation. **(c)** Initial SBI bias: limited variation over $[0, 2]$. **(d)** PTA-Aux duration: limited variation over $[0.1, 0.5]$, with degradation at 1.0, where the auxiliary remains active throughout most of training. Grey lines mark defaults; nonzero teal references are comparison levels rather than zero KL.

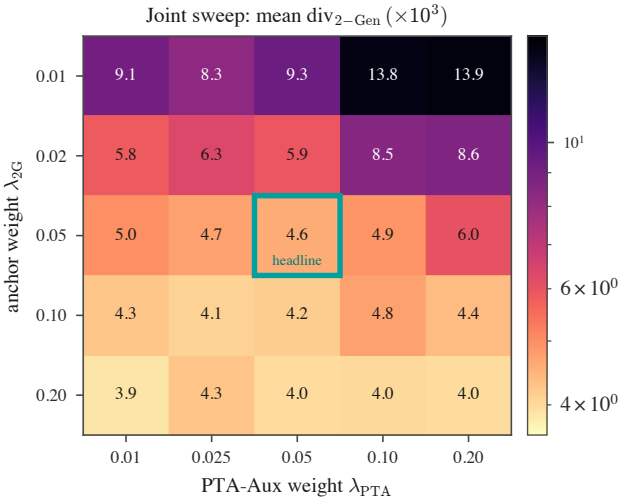


Figure 8: **Joint loss-weight sensitivity** at K=128. Values are mean $\mathrm{div}_{2\text{-Gen}} \times 10^3$ over a 5×5 grid of $(\lambda_{2G}, \lambda_{PTA})$. Larger anchor weights yield a broad region of low divergence, with smaller variation along the attention-loss axis. The outlined cell marks the main setting, λ_{2G}=λ_{PTA}=0.05. This grid evaluates predictive sensitivity within the displayed ranges.

D HYPERPARAMETER SENSITIVITY (FULL)

In addition to the K=128 panels, separate one-dimensional sweeps evaluate the loss coefficients at $K \in \{64, 512\}$ with two seeds per pool size. The results below summarize this four-setting evaluation.

Anchor weight sweep. Increasing λ_{2G} from 0.005 to 0.20 reduces mean divergence from 0.017 to 0.004. Improvements diminish beyond 0.05. This response is consistent with the anchor directly supervising the evaluated posterior target.

PTA-Aux weight sweep. Divergence is approximately 0.0065 over $\lambda_{\mathrm{PTA}} \in [0.01, 0.05]$ and increases at larger weights. One possible explanation is competition between attention regularization and next-token prediction during early training. Establishing that explanation would require measurements beyond the aggregate sensitivity curve. The selected value is 0.05.

Summary. The sweeps identify ranges with limited variation in predictive divergence, while also showing degradation outside those ranges. They support the selected configuration within this benchmark, without establishing invariance of phase boundaries or attention statistics to all coefficient changes.

E Order parameters: definitions and per-method values

We compute the order parameters of Gibson et al. (2026) from the end-of-training layer-one attention map $\mathbf{A}_1 \in \Delta_T^T$ at $K{=}128$, seed 0, on the same held-out batch used for divergences.

Previous-token attention weight $\varphi_\delta^{(1)} = \frac{1}{T-1}\sum_{i=1}^{T-1} A_1[i,\, i{-}1]$: the average mass assigned to the immediately preceding token. A value of 1.0 indicates all measured attention mass is on the preceding token. Under causal masking, the uniform-attention reference depends on the number of available positions.

Effective attended-positions count $n_A = \exp(H(A_1[T{-}1, \cdot]))$: the perplexity of the last-token attention distribution, i.e. the number of distinct past positions the model effectively averages over. Lower n_A indicates more concentrated attention; concentration alone does not establish the functional role of an attention head.

Kinetic phase boundary $K_1^* = \min\{K : \arg\min_\bullet \mathrm{KL}(\hat{\pi}^\bullet \| \pi^\theta) = \text{2-Gen}\}$ on the K-sweep: the smallest task-pool size at which the model is closest to $\hat{\pi}^{\text{2-Gen}}$ among the four Bayes predictors. Gibson et al. (2026) report $K_1^* \approx 94$ for the baseline.

Table 8: Order parameters per method (computed from $\mathbf{A}_1$ at $K{=}128$, seed 0). K_1^* from the full 18-cell grid.

Method	$\varphi_\delta^{(1)} \uparrow$	$n_A \downarrow$	$K_1^* \downarrow$
Two-Block baseline	0.40	22.6	64
Induction-Head Primer	0.40	44.4	64
Posterior Distillation	0.59	7.1	$\leq$32
SAIP (Ours)	**0.80**	**5.5**	$\leq$**32**

Comparison. Table 8 compares the measured attention statistics and phase boundaries. Previous-token attention is 0.40 for both the baseline and the Induction-Head Primer, 0.59 for Posterior Distillation, and 0.80 for SAIP. Thus the initial bias alone does not reproduce the full method's attention concentration. Posterior supervision can also accompany sharper attention without an explicit attention loss. SAIP has the lowest reported attention perplexity (5.5). The combined measurements distinguish concentrated attention from output accuracy and provide complementary descriptions of the final models.

F Per-method training dynamics and attention maps

Figure 9 reports the per-step convergence trajectory of $\mathrm{div}_{\text{2-Gen}}$ at $K{=}128$ for all four main methods, together with the end-of-training layer-one attention maps.

The additional five-seed trajectory summary gives final divergence $0.0046 \pm 1.1 \times 10^{-4}$ for SAIP and $0.033 \pm 5.8 \times 10^{-3}$ for the baseline after 8000 steps. The attention maps show more mass on the previous-token diagonal for SAIP and an intermediate pattern for Posterior Distillation.

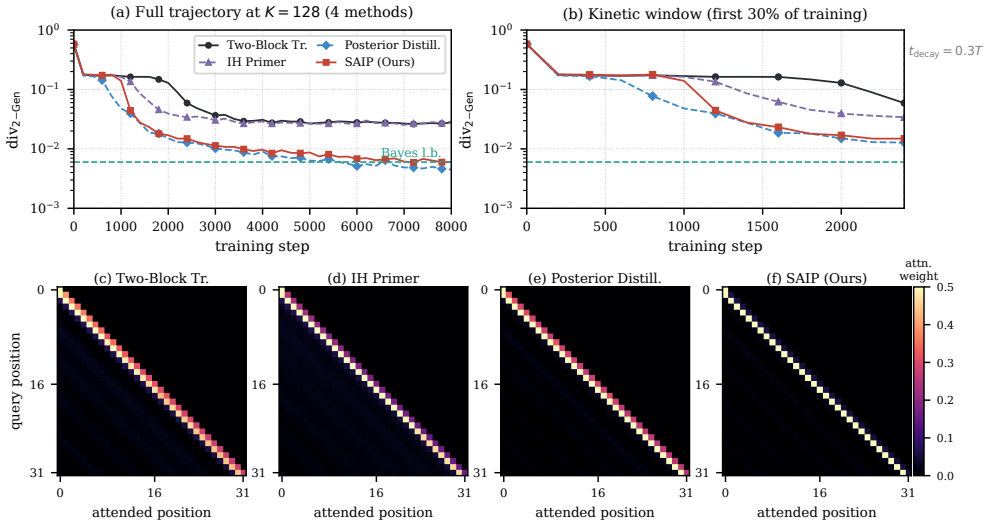


Figure 9: **Training dynamics and attention at K=128. (a)** Divergence over 8000 steps for four methods. SAIP and Posterior Distillation reach lower residual error than the cross-entropy baseline within the displayed budget. **(b)** Early-training trajectories over the interval in which PTA-Aux is active. After t_{decay}, SAIP optimizes cross-entropy plus the constant posterior anchor. **(c–f)** Final layer-one attention maps (a 32×32 patch). SAIP has the most concentrated previous-token diagonal among the displayed models.

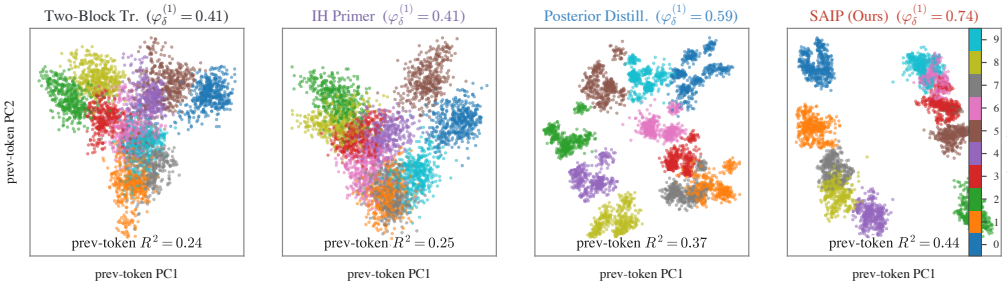


Figure 10: **Layer-one residual representations** (K=128, shared held-out batch). We project residuals onto the top two principal components of the per-x_{t-1} class centroids and color them by previous-token identity. The reported explained-variance statistic R^2 is $0.24/0.25$ for the cross-entropy methods, 0.37 for Posterior Distillation, and 0.44 for SAIP. The stronger separation by token identity accompanies higher previous-token attention.

These observations link the aggregate attention statistic to a visible change in attention structure. Convergence times should be compared within a common configuration and seed aggregation; different trajectory summaries are not interchangeable.

The residual projection in Figure 10 connects the attention measurements to the information represented after layer one. Previous-token identity explains more residual variation in SAIP, and the reported R^2 increases with $\varphi_\delta^{(1)}$ across the four methods. This association is consistent with the previous-token circuit interpretation.

Figure 11 records the timing of attention formation. SAIP exhibits the initialized diagonal at step zero and a more concentrated pattern within approximately 500 steps, before the auxiliary ends at step 2400. The baseline develops the pattern later and retains more off-diagonal mass at the end of training. Persistence after PTA-Aux ends shows that the trained pattern remains under cross-entropy and the posterior anchor; it does not isolate which remaining term maintains it.

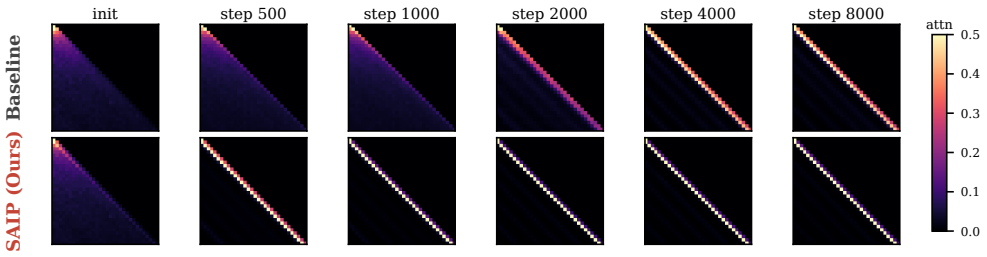


Figure 11: **Previous-token attention over training** (layer-one 32×32 patches; queries in rows, attended positions in columns). The baseline develops a more diffuse diagonal over 8000 steps. SAIP begins with the initial bias and develops a concentrated diagonal within the first 500 steps. The pattern remains visible after PTA-Aux ends at t_{decay}=2400.

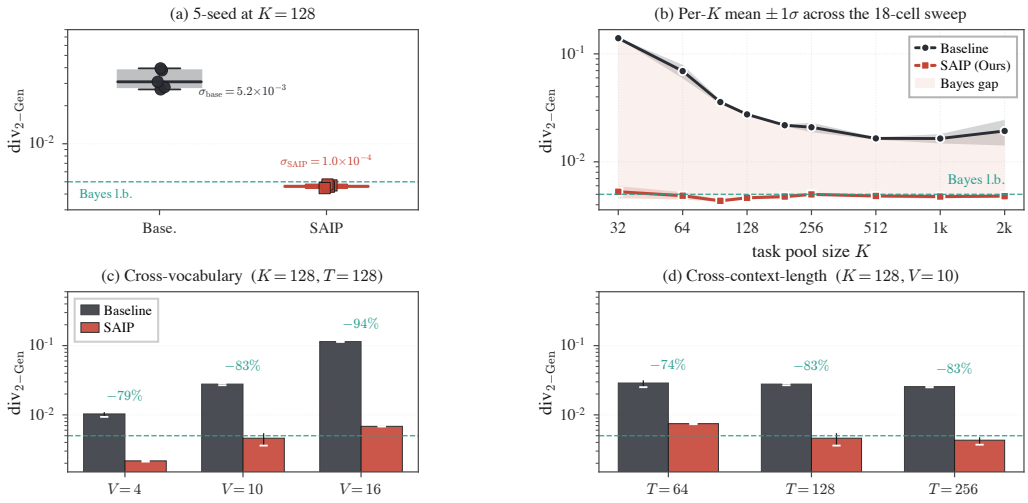


Figure 12: **Seed variability and transfer across configurations.** **(a)** Five-seed distributions at K=128, V=10, T=128; each marker denotes one seed. Standard deviations are 1.0×10^{-4} for SAIP and 5.2×10^{-3} for the baseline. **(b)** Per-K means and standard-deviation bands for the 18-setting sweep; shading marks divergence reduction. **(c)** Vocabulary variation gives 79–94% reduction. **(d)** Context-length variation gives 74–83% reduction. Coefficients are held fixed.

Figure 12 evaluates seed variability and changes in task configuration. In the five-seed comparison, the reported standard deviation is approximately $50\times$ lower for SAIP than for the baseline. Across the tested vocabulary sizes and context lengths, relative divergence reductions remain 74–94% without changing coefficients. The combination of seed repetitions and configuration sweeps evaluates sensitivity to both initialization and task specification.

All 18 evaluated settings improve over the baseline and exceed the specified 20% reduction threshold (Figure 13a–c). Thus the aggregate improvement is not confined to a subset of the sampled settings. The final attention-map difference in panel (e) is concentrated on the previous-token diagonal. This spatial correspondence links the aggregate attention statistic to the previous-token pattern targeted by the auxiliary loss.

Figure 14 compares methods at the ablation coefficient λ_{2G}=0.02. The baseline divergence and the relative improvement vary with task-pool size. These values should be distinguished from the main 0.05 evaluation. A divergence heatmap alone does not identify the nearest of all four Bayes predictors, so strategy assignments are assessed separately in Figure 2.

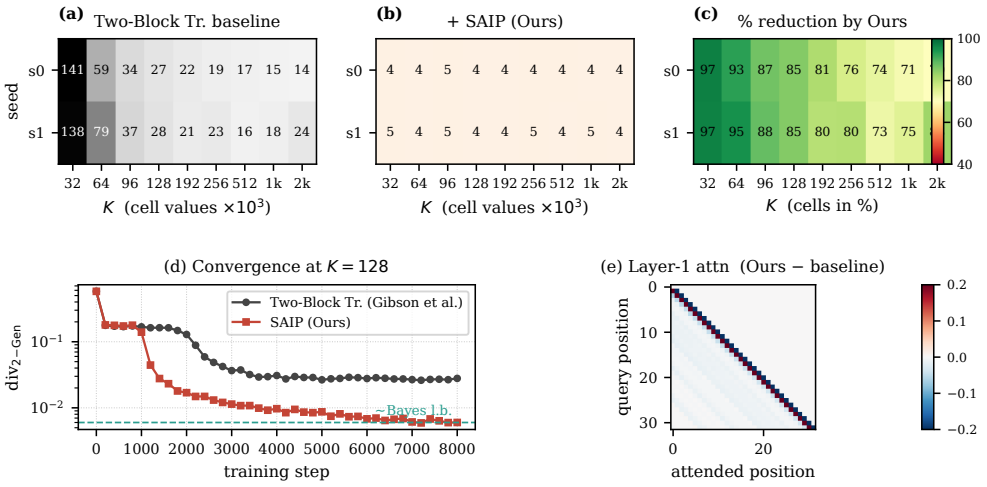


Figure 13: **Per-cell and per-step detail. (a–c)** Per-cell heatmap of $\text{div}_{\text{2-Gen}}$ for each $K \times$ seed cell. (a) Two-Block Transformer baseline. (b) SAIP (Ours, $\lambda_{2G}{=}0.05$). Cell values are $\times 10^3$. (c) % reduction by Ours; all 18 cells exceed the 20% threshold. **(d)** Training trajectory at $K{=}128$. SAIP crosses an order-of-magnitude reduction by step$\approx$3500; the baseline plateaus an order of magnitude higher. **(e)** Difference between the layer-one attention maps of Ours and the baseline at end-of-training (red = Ours larger). The previous-token diagonal that PTA-Aux targets is visibly enhanced.

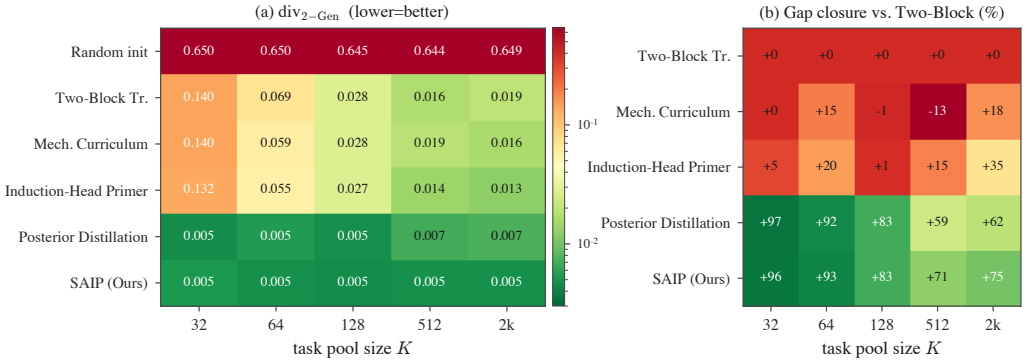


Figure 14: **Comprehensive method$\times K$ comparison** (illustrative 2-seed run at the $\lambda_{2G}{=}0.02$ ablation default; main 5-seed $\lambda_{2G}{=}0.05$ numbers in Table 1). **(a)** Mean $\text{div}_{\text{2-Gen}}$ for six methods at five representative K values straddling the $K_1^*{\approx}94$ boundary (log-colored). **(b)** Per-cell gap closure vs. the Two-Block baseline (positive = better than baseline). SAIP closes a large fraction of the gap at every K; Posterior Distillation is competitive at small K but degrades at $K{\geq}512$ where the baseline is closer to optimal; Mech. Curriculum and IH Primer have divergences closer to the baseline.

Figure 15 compares divergence to all four posterior references, distinguishing the targeted 2-Gen improvement from the smaller changes in the other metrics. Figure 16 complements this comparison with attention concentration, phase assignment, and cross-entropy retention; these attributes describe different properties of the trained models.

G Computational Cost and the Cost–Quality Frontier

The attention-based methods were timed on a shared RTX 4090 using the common training implementation. Table 9 reports median time per (K, seed) setting across the timing grid, alongside mean posterior divergence. Medians reduce sensitivity to occasional contention-related timing

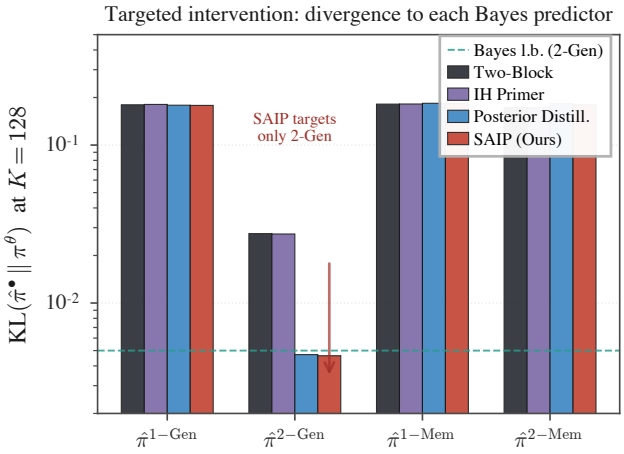


Figure 15: **Divergence to the four Bayes predictors** at $K{=}128$ (2-seed evaluation at the 0.02 ablation coefficient). The largest between-method change is in $\mathrm{div}_{2\text{-Gen}}$; the other posterior divergences vary less. Main-setting results are reported separately in Table 3.

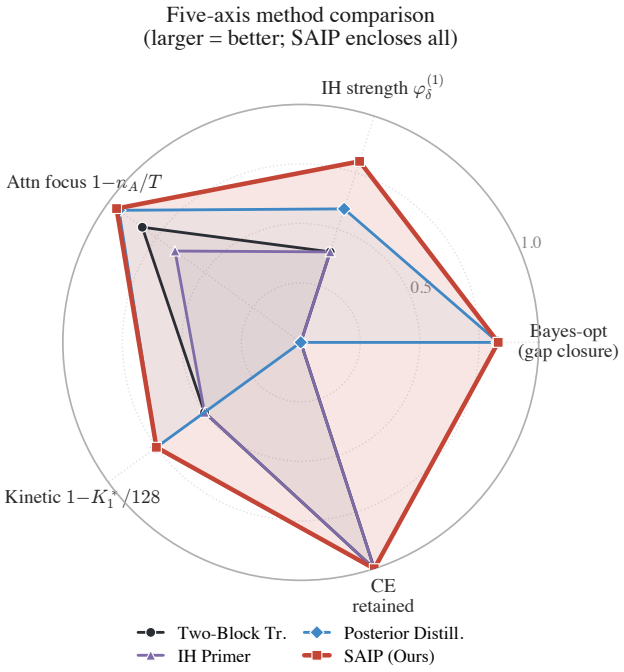


Figure 16: **Comparison across five reported attributes** at $K{=}128$. Axes represent relative divergence reduction, previous-token attention $\varphi_\delta^{(1)}$, attention concentration $1 - n_A/T$, the rescaled phase boundary $1 - K_1^*/128$, and retention of cross-entropy. SAIP combines low divergence with concentrated attention and cross-entropy supervision. The axes summarize distinct attributes rather than a common scalar objective.

outliers. Mamba and CTW are not included in this timing comparison. Figure 4c displays the resulting time–divergence measurements.

SAIP has a median measured time of 77.6 s, compared with 78.4 s for the Two-Block baseline and 81.5 s for Posterior Distillation. Among the listed two-block methods, it has the lowest reported mean

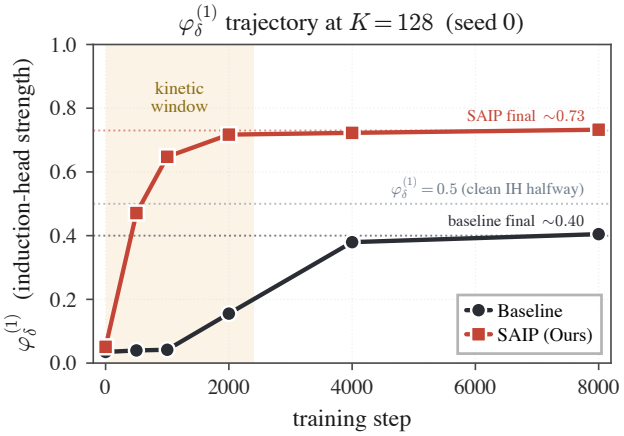


Figure 17: **Previous-token attention during training** (K=128, seed 0; six checkpoints). SAIP exceeds $\varphi_\delta^{(1)}$=0.5 while PTA-Aux is active and maintains high attention after t_{decay}=2400. The baseline approaches approximately 0.40 later in training. The comparison describes attention concentration rather than a binary criterion for circuit formation.

Table 9: **Measured training time and posterior divergence.** Median time per setting on a shared RTX 4090 and mean $\text{div}_{\text{2-Gen}}$ for the timing evaluation. SAIP and the Two-Block baseline have similar median times. Timing-grid means need not coincide with the differently aggregated method comparison in Table 1.

Method	median s/cell	mean $\text{div}_{\text{2-Gen}}$
Single-Block Tr.	**57.6**	0.160
SAIP (Ours)	77.6	**0.005**
Two-Block Tr. baseline	78.4	0.054
Induction-Head Primer	79.7	0.048
Posterior Distillation	81.5	0.007
Mechanism Curriculum	90.0	0.052

divergence and median time. The Single-Block model is faster (57.6 s) but has higher divergence. These point estimates describe the observed trade-off; repeated paired timings would be needed to establish runtime equivalence or a reliable speed advantage. The auxiliary losses involve additional operations even when the measured overhead is small.

H In-context dynamics and phase structure

Figure 18 evaluates $\text{div}_{\text{2-Gen}}$ at each context position, rather than at each training step. SAIP and Posterior Distillation have lower divergence across the displayed positions than the cross-entropy baseline. Figure 2 provides a complementary measurement: the nearest of the four Bayes predictors over task-pool size and training step. This is a strategy classification based on output divergence, not a direct measurement of an optimization basin. SAIP enters the 2-Gen regime at the smallest tested pool, giving an empirical boundary no greater than 32. The discrete estimate and the approximately 94 reference value from Gibson et al. (2026) arise from different evaluations and should not be treated as identical measurements.

The position-wise evaluation shows that the prediction advantage extends beyond the final-token metric. The phase map additionally shows changes in nearest-predictor assignment at smaller pools. These observations should be interpreted separately: where the baseline already belongs to the 2-Gen regime, reduced divergence represents a more accurate approximation within that regime. The full improvement therefore cannot be attributed exclusively to a categorical phase change.

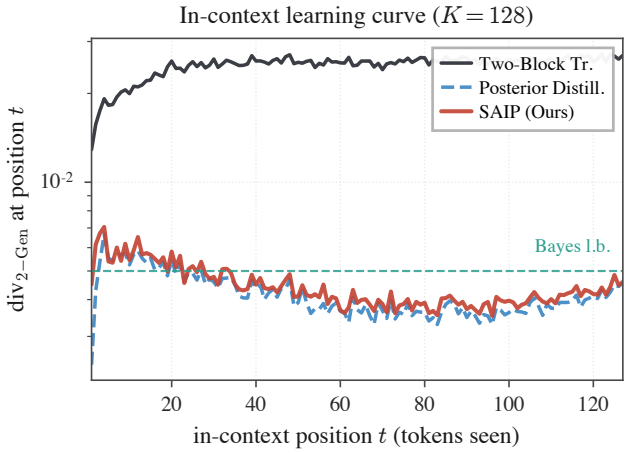


Figure 18: **Position-wise posterior divergence** at K=128 on a held-out batch of 512 sequences. SAIP and Posterior Distillation have lower divergence than the cross-entropy baseline across the displayed positions. This evaluates the accuracy of context-dependent predictions beyond the final token.

Order-parameter dependence on task-pool size. Figure 19 treats K as a control parameter and compares final attention statistics. For the cross-entropy methods, $\varphi_\delta^{(1)}$ increases from approximately 0.18 at K=32 to 0.46 at K=2048. SAIP has a higher, comparatively flat profile near 0.72 in this evaluation, together with lower posterior divergence. The plotted profiles suggest different responses to task diversity under the training procedures. They do not establish a dynamical bifurcation in the mathematical sense, which would require an analysis of stability or limiting dynamics.

Recovered transition matrix. We additionally compare recovered transition matrices. Predictions from 512 sequences generated by a held-out matrix M^* are grouped by the current token at positions $t \geq 48$, producing a $V \times V$ recovered matrix (Figure 20). The finite-context 2-Gen posterior provides the reference prediction. SAIP has row-wise L_1 distance 0.035 to this reference, compared with 0.116 for the baseline, approximately 3.3× larger. This probe evaluates agreement with the posterior at the level of aggregated transition predictions.

To evaluate variability across tasks, we repeat the probe over 16 independently sampled held-out chains (Table 10). Mean row-wise distances to the posterior estimate are 0.030 ± 0.002 for SAIP, 0.029 ± 0.002 for Posterior Distillation, and 0.105 ± 0.012 for the baseline. The pattern therefore persists across these sampled chains. The posterior estimate itself has distance 0.389 ± 0.020 to M^*, compared with 0.397 ± 0.022 for SAIP. This posterior-to-matrix distance is an empirical reference, not a proven L_1 lower bound for arbitrary estimators or individual chains. Bayesian optimality for predictive log loss does not establish such an L_1 bound.

Table 10: Recovered-matrix distances at K=128 (row-wise L_1, mean±s.d. over 16 held-out chains). Distance to the posterior measures agreement with the Bayesian reference; distance to M^* measures agreement with the generating matrix. These are distinct evaluation criteria.

Model	L_1 to 2-Gen Bayes	L_1 to M^*
$\hat{\pi}^{\text{2-Gen}}$ Bayes (target)	0.000	$0.389_{\pm.020}$
SAIP (Ours)	$\mathbf{0.030}_{\pm.002}$	$0.397_{\pm.022}$
Posterior Distillation	$0.029_{\pm.002}$	$0.389_{\pm.021}$
Two-Block baseline	$0.105_{\pm.012}$	$0.429_{\pm.032}$

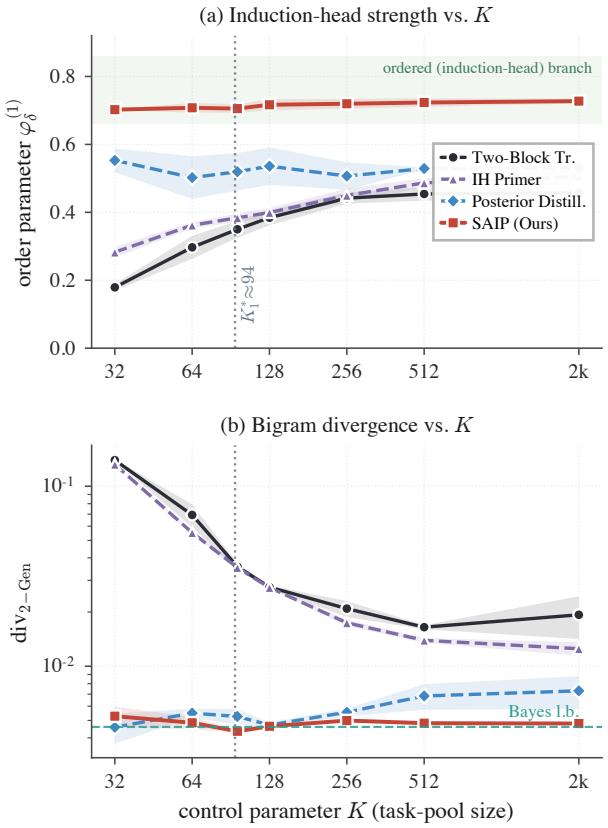


Figure 19: **Attention and divergence across task-pool size** (2-seed mean±s.d.). **(a)** Previous-token attention increases with K for the cross-entropy methods, while SAIP has a higher profile near 0.72. Posterior Distillation is intermediate. **(b)** Divergence decreases with K for the baselines; SAIP has lower residual error over the sweep. The reference boundary $K_1^* \approx 94$ is included for comparison.

I Parameter interpolation and depth-wise prediction probes

We characterize the parameter configurations using linear interpolation (Frankle et al., 2020). Baseline, Posterior Distillation, and SAIP models use the same $K{=}128$ task pool. We evaluate $\mathrm{div}_{2\text{-Gen}}$ along $\theta(a) = (1-a)\theta_X + a\theta_Y$ for each pair. The baseline embeds into the SAIP parameterization at $\alpha_{\mathrm{SBI}}{=}0$; the corresponding forward outputs were checked to agree before interpolation. Each tested path has an interior divergence larger than either endpoint. For baseline–SAIP, the maximum is 0.074, compared with endpoint values 0.027 and 0.005 (Table 12). This establishes a barrier along the evaluated straight path, not disconnected optimization basins or the absence of a low-divergence nonlinear path.

The logit-lens probe decodes residual states using the final LayerNorm and tied output head at three depths (Figure 21b). After layer one, the methods have divergences of approximately 0.39–0.45. The larger performance difference appears at the final read-out: the baseline gives 0.027, compared with approximately 0.004–0.005 for the posterior-supervised methods. Thus the difference is most visible in the output decoded after layer two. Because the decoder is shared across depths, the probe measures accessible predictions under that decoder rather than all information represented at intermediate layers.

Exact values for both panels of Figure 21 are tabulated in Section N.

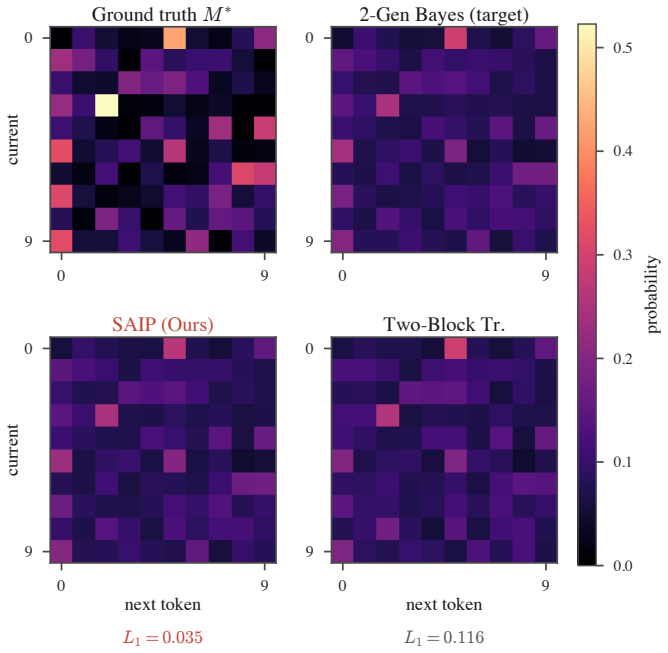


Figure 20: **Recovered transition predictions** for one held-out matrix M^* (K=128, V=10). Predictions are grouped by current token over context positions $t \geq 48$. Row-wise L_1 distances to the 2-Gen posterior estimate are 0.035 for SAIP and 0.116 for the baseline. The posterior estimate and the generating matrix differ because the context contains finite evidence.

Table 11: Mamba (our pure-PyTorch SSM reimplementation), 2 seeds/cell. Unigram divergence decreases with K, while substantial bigram-posterior divergence remains under this configuration.

	metric	32	64	128	512	2048
1-layer	$\text{div}_{\text{2-Gen}}$	0.135	0.130	0.136	0.134	0.135
	$\text{div}_{\text{1-Gen}}$	0.189	0.135	0.101	0.070	0.060
2-layer	$\text{div}_{\text{2-Gen}}$	0.138	0.124	0.119	0.122	0.132
	$\text{div}_{\text{1-Gen}}$	0.285	0.196	0.136	0.087	0.063

J Mamba (state-space) baseline details

Table 11 reports results for our compact PyTorch selective-SSM implementation (d=64, d_{state}=16, causal convolution width 4, expansion factor 2). The implementation uses standard PyTorch operations rather than an optimized selective state-space kernel. Both one- and two-layer versions reduce unigram divergence as K increases, but retain bigram divergence near 0.13. We treat them as implementation-specific comparisons, distinct from theoretical analyses such as Bondaschi et al. (2026).

K Relationship to additional learning settings

The following studies address related architectures or task distributions. Their relationship to our benchmark motivates comparisons but does not make our implementations exact reproductions of their full experimental settings.

Selective induction heads (D'Angelo et al., 2025). Selective-lag settings require adaptation to dependencies beyond the fixed first-order transition used here. Our two-block baseline evaluates the fixed-lag setting; it should not be interpreted as a reproduction of the broader selective-lag construction.

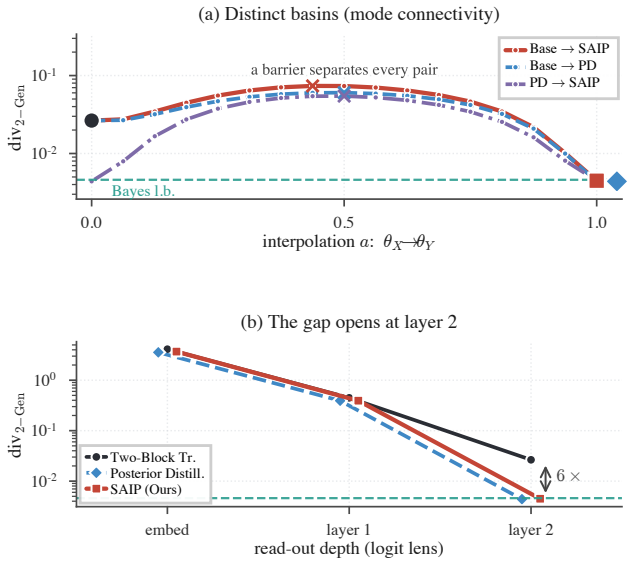


Figure 21: **Parameter interpolation and logit-lens probes** (K=128). **(a)** Divergence along straight parameter paths between trained models. Each path has an interior barrier, including the Posterior Distillation–SAIP path. This does not establish topological separation. **(b)** Posterior divergence decoded at three depths. Method differences are largest at the final read-out. Exact values are in Appendix N.

Depth and representational capacity (Ekbote et al., 2026). Theoretical depth analyses motivate the Single-Block and Two-Block comparison. Our finite-width, finite-training results describe empirical performance in the chosen setting and do not independently prove a representational lower bound.

Variable-order Markov learning (Zhou et al., 2024). We include order-1 CTW with KT smoothing as a classical comparison. Its reported divergence is 0.020. The KT parameter α=0.5 differs from the α=1 posterior reference, which is a potential contributor to this discrepancy; the experiment does not isolate it as the sole cause.

State-space prediction (Bondaschi et al., 2026). Mamba provides an architectural comparison beyond attention. Our implementation-specific results and their limitations are reported in Appendix J.

Markov pre-training for recommendation (Xu et al., 2026). Downstream recommendation evaluates a different objective and data distribution from the controlled posterior divergence studied here. We therefore discuss this application as related context rather than as a directly comparable experimental baseline.

Dynamics beyond the Markov benchmark. Mainali & Teixeira (2025) derive learning dynamics for linear transformers and use spectral and subspace measurements to investigate nonlinear models. Their dynamical analysis differs from our static interpolation probe, which measures divergence along selected parameter paths.

Task dependence of learning modes. Ku et al. (2026) relate ICL/IWL preferences to environmental stability, cue reliability, and optimization costs in regression and classification. This supports treating task statistics as part of the explanation for a learned strategy rather than assuming a universal transition threshold.

Temporal retrieval in language models. Bajaj et al. (2026) use head ablations to connect induction heads with temporal retrieval and serial recall. Their task-specific causal measurements complement our descriptive attention probes, but do not establish the causal role of the heads in our Markov models.

L PSEUDOCODE

Algorithm 1 SAIP training step at step t (Python-like pseudocode)

Require: batch $x \in \{1, \dots, V\}^{B \times T}$, model with learnable α_{SBI}, schedules $\lambda_{\mathrm{PTA}}(t)$, $\lambda_{2\mathrm{G}}$

logits $\leftarrow \mathrm{model}(x)$ {layer-one attn adds $\alpha_{\mathrm{SBI}} \cdot \mathbf{M}_{\mathrm{prev}}$ before softmax}
$\mathcal{L}_{\mathrm{CE}} \leftarrow \mathrm{CrossEntropy}(\mathit{logits}, x_{\text{shifted}})$
if $t < t_{\mathrm{decay}}$ **then**
  $\mathbf{A}_1 \leftarrow$ realized layer-one attention (recomputed with grad)
  $w \leftarrow \frac{1}{2}\lambda_{\mathrm{PTA}}(1 + \cos\frac{\pi t}{t_{\mathrm{decay}}})$
  $\mathcal{L}_{\mathrm{PTA}} \leftarrow w \cdot \mathrm{KL}(\mathbf{M}_{\mathrm{prev}} \,\|\, \mathbf{A}_1)$
else
  $\mathcal{L}_{\mathrm{PTA}} \leftarrow 0$
end if
$\hat{\pi}^{\text{2-Gen}} \leftarrow \mathrm{ClosedForm2Gen}(x; V, \alpha)$ {vectorized, no grad}
$\mathcal{L}_{2\mathrm{G}} \leftarrow \lambda_{2\mathrm{G}} \cdot \mathrm{KL}(\hat{\pi}^{\text{2-Gen}} \,\|\, \mathrm{softmax}(\mathit{logits}))$ {forward KL with oracle target}
$\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{PTA}} + \mathcal{L}_{2\mathrm{G}}$
optimizer.step($\mathcal{L}$)

The closed-form 2-Gen predictor is computed once per batch from running bigram counts of the input sequence and is fully vectorized (no in-graph dependence on the model). The PTA-Aux term recomputes the layer-one attention with a fresh forward pass on the same batch, so the gradient flows through both the attention logits and α_{SBI}.

M REPRODUCIBILITY CHECKLIST

- Implementation: the trainable methods use a shared PyTorch implementation for model construction, optimization, and evaluation, with method-specific objectives and curriculum schedules specified in Appendix B.
- Compute: experiments use an RTX 4090 GPU. The median training time is approximately 78 s per baseline or SAIP setting, giving approximately 23 min for an 18-setting sweep on the measured shared GPU. The anchor and attention auxiliary require additional operations; timings are reported in Appendix G.
- Randomness: sampling uses explicitly seeded random-number generators. Deterministic execution is requested where supported, while unsupported operations issue warnings. Some GPU cumulative-sum operations remain nondeterministic, so bitwise reproducibility is not guaranteed.
- Closed-form Bayes predictors: the 2-Gen target $\hat{\pi}^{\text{2-Gen}}$ is computed in vectorized form from cumulative bigram counts, with complexity $O(B\,T\,V^2)$.
- Evaluation: same held-out batch of 512 length-T sequences across all methods at the same (K, seed) cell.

N NUMERICAL VALUES FOR THE LANDSCAPE AND LENS PROBES

The following tables give the measured values for Figure 21 at $K{=}128$, using a shared task pool and a 512-sequence evaluation batch. The baseline embedding used for interpolation was checked numerically, with $\max|\Delta\mathrm{logit}| = 0$ in that check.

Table 12 reports the interpolation measurements, while Table 13 and Figure 22 report predictions decoded at each depth. The Posterior Distillation–SAIP interpolation reaches divergence 0.055, above its endpoints 0.004 and 0.005. Thus similarly accurate predictors need not be connected by a low-divergence straight parameter path. The depth-wise probe additionally shows that the method differences are largest at the final read-out. Together, the probes describe distinctions between trained parameter configurations and decoded predictions, without establishing a unique causal explanation for them.

Table 12: **Linear-mode-connectivity barriers** between trained solutions: $\text{div}_{\text{2-Gen}}$ at the two endpoints, the maximum along the straight parameter path, and its location a^*.

Pair $X \to Y$	end X	end Y	barrier	a^*
Baseline $\to$ SAIP	0.027	0.005	0.074	0.44
Baseline $\to$ Post. Distill.	0.027	0.004	0.060	0.50
Post. Distill. $\to$ SAIP	0.004	0.005	0.055	0.50

Table 13: **Logit-lens read-outs**: $\text{div}_{\text{2-Gen}}$ of the prediction read from the residual stream (through the final LayerNorm and tied head) at three depths. The methods agree through layer 1; the baseline/SAIP split appears at the layer-2 read-out.

Method	embedding	after layer 1	after layer 2
Two-Block Tr.	4.12	0.447	0.027
Post. Distill.	3.56	0.392	**0.004**
SAIP (Ours)	3.69	0.393	**0.005**

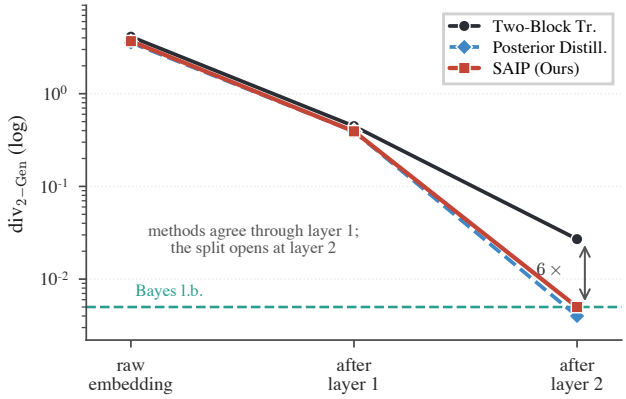


Figure 22: **Logit-lens measurements by depth** (Table 13). Decoding with the final LayerNorm and tied head gives similar layer-one divergences for the three methods. Final-layer divergence is 0.027 for the baseline and 0.004–0.005 for the posterior-supervised methods.

Scope of the probes. Interpolation evaluates straight paths at one $(K{=}128, \text{seed } 0)$ setting; nonlinear connectivity and variability across seeds are not measured. The logit lens uses the final LayerNorm and tied head at intermediate depths, so it measures predictions accessible to that decoder. Identical probe procedures support comparison across the evaluated models, but do not remove these interpretive limitations.