# SPATIAL INDUCTION HEADS: IN-CONTEXT LEARNING OF MULTIDIMENSIONAL CELLULAR AUTOMATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Induction heads provide a mechanistic account of in-context learning in sequential data, but existing theory largely assumes that the context relevant to a prediction forms a contiguous block. In multidimensional data, serialization breaks this assumption by scattering spatial neighbors across distant positions in the token sequence. We study how transformers overcome this routing problem in multi-dimensional stochastic and deterministic cellular automata, where each trajectory is generated by an unknown local rule and presented as a flattened sequence without an explicit coordinate-based spatial inductive bias. We introduce *spatial induction heads*, two-layer gather-and-match circuits in which the first layer reconstructs the relevant spatial neighborhood and the second matches the resulting configuration against earlier occurrences. We give two explicit realizations of the gather and show that the positional dimension required for spatial routing depends only on the local neighborhood and spatial dimension, not on grid volume or trajectory horizon. We further construct a matching layer which implements Bayesian counting. The end-to-end circuit can approximate the Bayesian posterior arbitrarily closely for stochastic rules and can predict exactly for deterministic rules. Empirically, trained two-layer transformers generalize to unseen rules in one and two dimensional settings, achieving near-perfect deterministic rollouts and less than $0.005$ nats KL from the Bayes-optimal predictor on stochastic rules. Attention patterns and layerwise probes align with the predicted gather-and-match computation, providing mechanistic evidence for spatial induction in trained transformers.

## 1 INTRODUCTION

The *induction head* provides one of the best-understood mechanistic accounts of in-context learning (Elhage et al., 2021; Olsson et al., 2022). This two-layer circuit predicts novel sequences by completing patterns from precedent: the first layer records preceding tokens, while the second layer matches the current context to past records and copies the subsequent tokens. Induction heads have been used to explain the in-context behavior of large language models (Brown et al., 2020; Wei et al., 2022), and their theoretical foundations have been rigorously established for Markov chains (Bietti et al., 2023; Edelman et al., 2024). In this setting, the *statistical induction head* attends to the preceding $k$ tokens to retrieve the empirical transition law (Ekbote et al., 2026). However, this theory relies on a rarely emphasized assumption: the context block is a contiguous sequence immediately preceding the query, reachable by a single attention layer using relative position.

When transformers are applied to multidimensional spatiotemporal data, this assumption breaks down. Serializing a spatial grid into a one-dimensional token stream destroys local adjacency, scattering immediate neighbors across distant intervals in the sequence. The original Vision Transformer (Dosovitskiy, 2020) uses learned 1D absolute positional embeddings and shows that these embeddings recover aspects of the underlying 2D grid structure. Learned absolute positional embeddings also appear in recent image generators (Li et al., 2024; Yu et al., 2025; Bachmann et al., 2025). Their success motivates studying how Transformers learn structural relationships from token sequences without explicit geometric inductive biases on their positional embeddings.

Despite this empirical success, there is no mechanistic understanding of how a transformer leverages a learned positional table to gather a neighborhood fractured by serialization, nor how it performs *spatial in-context learning*. Existing theories for $k$-order Markov chains assume contiguous context



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.

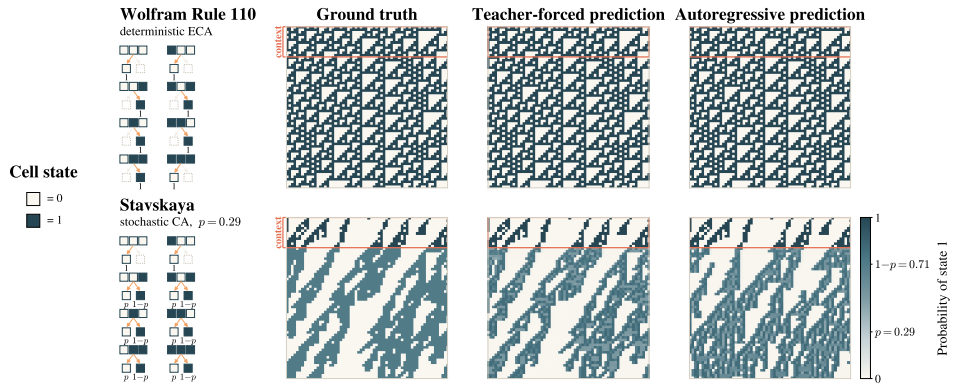


Figure 1: **Trained-model predictions on two held-out rules.** Top row: Wolfram Rule 110, a Turing-complete deterministic ECA. Bottom row: Stavskaya, a stochastic CA with $p$=0.29; predictions are colored by the probability the model assigns to state 1. Teacher-forced predictions (middle column) condition on the ground-truth history at every step; autoregressive predictions (right column) are rolled out from the context alone, conditioning on the sample from the previous timestep.

(Edelman et al., 2024), offering no analogue for sequential-to-spatial routing. This raises two fundamental questions: How does a transformer address and assemble scattered tokens to reconstruct a spatial neighborhood? And once assembled, how does it optimally infer the underlying dynamical rule without any coordinate-based spatial inductive bias?

We address these questions by studying transformers on multidimensional stochastic cellular automata, which are historically viewed as systems of *locally interacting Markov chains* evolving on a multidimensional grid (Fernández et al., 2018). A stochastic cellular automaton consists of a grid of finite-state cells updated in parallel via a local kernel, mapping a cell's neighborhood at one step to a state distribution at the next (Definition 1). In our study, the kernel is sampled afresh for each trajectory from a Dirichlet prior and hidden from the model, requiring pure in-context inference. This provides an ideal testbed: simple local rules generate essentially unbounded complexity (Langton, 1990), and the dependence structure is a spatial analogue to the Markov setting. We train two-layer transformers on this task. Figure 1 shows them continuing a trajectory from a short context on two held-out rules, the deterministic Wolfram Rule 110 and the stochastic Stavskaya rule at $p = 0.29$.

**Contributions.** We formalize spatial in-context learning as a problem where serialization hides the geometry that the model must exploit. Our contributions are as follows:

- **Task & Theoretical Framework:** We consider multidimensional cellular automata and derive the closed-form Bayes-optimal predictor (Theorem 3). We then introduce the *spatial induction head* (Definition 5), a two-layer circuit whose first layer *gathers* the spatial dependencies needed at each token, and whose second *matches* the resulting configuration against precedent.
- **Mechanistic Constructions:** We give two explicit realizations of the gather, one using a head per offset and one using a single content-routed head with a separating MLP (Theorem 7, parts (I) and (II)). Both route scattered spatial neighbors with a compact set of positional frequencies whose number is independent of grid volume and horizon (Lemma 6).
- **Bayesian Matching:** We show the matching layer performs Bayesian counting: once the gathered representation is available, a single Layer-2 head computes the posterior exactly under hard attention (Proposition 8) and arbitrarily closely otherwise (Theorem 9).
- **Empirical Verification:** We validate the theory on 1D and 2D cellular automata. Two-layer transformers generalize to held-out rules (Figure 1), surpassing prior ICL results on cellular automata (Section 2), while attention maps (Figures 2 and 3) and layerwise probes (Figure 4) give strong evidence for the predicted gather-and-match computation.

## 2 Related work

**Transformers for automata and dynamical systems.** Transformers have been studied on discrete automata both as diagnostic tasks and as synthetic sequence models. Vafa et al. (2024) use determin-

istic finite automata to probe limitations of large models, while Liu et al. (2022); Berkovich et al. (2025); Burtsev (2024) train smaller transformers on cellular automata. Only Burtsev (2024) studies in-context generalization to unseen rules; Liu et al. (2022) train per automaton and Berkovich et al. (2025) provide the rule in the prompt. Related work finds in-context prediction in discretized chaotic systems and hidden Markov processes (Liu et al., 2024; Dai et al., 2026), but does not identify a circuit for recovering spatially scattered dependencies.

**Induction heads and Markov-chain ICL.** Induction heads originate in mechanistic studies of language models (Elhage et al., 2021; Olsson et al., 2022; Conmy et al., 2023). Theory has since characterized their operation on $k$-order Markov chains (Bietti et al., 2023; Makkuva et al., 2024a;b; Nichani et al., 2024; Rajaraman et al., 2024), including statistical induction heads (Edelman et al., 2024) and a two-layer construction for general $k$ (Ekbote et al., 2026). Our setting extends this line: probabilistic cellular automata are precisely systems of locally interacting Markov chains (Fernández et al., 2018). Serialization scatters the $k$ relevant neighbors through the token stream, so the model must first gather them before performing the usual induction head computation.

**Bayesian and algorithmic views of in-context learning.** A complementary account interprets ICL as approximate posterior inference under a prior induced by pretraining (Xie et al., 2021; Müller et al., 2021). In our stochastic cellular-automaton setting, the Bayes predictor is a smoothed count, and we show that the matching head represents it. Another line shows that transformers can implement optimization procedures such as gradient descent and higher-order variants (Von Oswald et al., 2023; Zhang et al., 2024; Ahn et al., 2023; Giannou et al., 2024); Cole et al. (2025) extends this view to linear dynamical systems. These works characterize an implemented learning algorithm, without considering spatial structure, whereas we characterize a circuit and the estimator it computes.

**Spatial structure in transformers.** Architectures for grid-structured data often supply geometry explicitly, as in weather models using coordinate-dependent attention or mesh graphs (Bi et al., 2023; Lam et al., 2023). Vision transformers instead flatten patches and can use learned one-dimensional absolute position embeddings, with learned embeddings recovering row–column structure (Dosovitskiy, 2020). We adopt this geometry-free setting and ask how a standard transformer uses learned positions to route a scattered neighborhood and perform spatial in-context inference.

## 3 The task and the routing problem

In this section, we cover four key pieces of background. First, we need well specified geometric data, so we consider the setting of stochastic cellular automata (§3.1). Then, we describe how we flatten the grid into a one-dimensional token stream, and we identify the set of offsets each prediction must reach (§3.2). Furthermore, we describe the two-layer transformer we study, which is given no geometric prior beyond a learned absolute positional table (§3.3). Finally, we identify the queries on which spatial in-context learning is well posed, and give the Bayes-optimal predictor for them (§3.4).

### 3.1 Cellular automata

**Definition 1** (Stochastic cellular automaton)**.** Fix a torus $\mathcal{G} = [W_1] \times \cdots \times [W_d]$, writing $[n] \coloneqq \{0, \ldots, n-1\}$, of volume $|\mathcal{G}| \coloneqq \prod_{j=1}^{d} W_j$, with $W_j \geq 2$. Cells of $\mathcal{G}$ are written $z = (z_1, \ldots, z_d)$. Let $\mathcal{S} = [V]$ ($V \geq 2$) be a state set and $\delta_1, \ldots, \delta_k \in \mathbb{Z}^d$ a *neighborhood* of $k \geq 1$ offsets, distinct modulo $(W_1, \ldots, W_d)$. A *local kernel* $f : \mathcal{S}^k \to \Delta^{V-1}$, with $\Delta^{V-1}$ the probability simplex on $\mathcal{S}$, defines a trajectory over horizon $T \geq 2$ from initial states $(x_{0,z})_{z \in \mathcal{G}}$, drawn i.i.d. uniformly on $\mathcal{S}$, via transitions for $t \in \{1, \ldots, T-1\}$, conditionally independent across $z$ given step $t-1$:

$$x_{t,z} \;\sim\; f(\cdot \mid P_{t,z}), \qquad P_{t,z} \;\coloneqq\; \big(x_{t-1,\, z+\delta_1},\; \ldots,\; x_{t-1,\, z+\delta_k}\big) \in \mathcal{S}^k,$$

where addition is modulo $(W_1, \ldots, W_d)$. We write $(t, z)$ for the cell $z$ at step $t$. $P_{t,z}$ is its *parent configuration*, whose entries are ordered by $\delta_1, \ldots, \delta_k$, and we write $D_K \coloneqq \{\delta_1, \ldots, \delta_k\}$ for the set of *key offsets*, so $|D_K| = k$. The kernel $f$ is drawn from the prior $\bigotimes_{c \in \mathcal{S}^k} \mathrm{Dir}(\alpha \mathbf{1}_V)$ with $\alpha > 0$ and $\mathbf{1}_V$ the all-ones vector, so the rows $f(\cdot \mid c)$, one per configuration $c$, are independent.

This is a standard probabilistic cellular automaton on a torus (Fernández et al., 2018). The neighborhood $D_K$ is fixed, and we ask whether the model can recover it. The kernel $f$ is drawn afresh for every trajectory and is never shown to the model, whose parameters are fixed at test time. It observes

only the flattened trajectory, and must infer $f$ from earlier occurrences of the same neighborhood configuration within that one trajectory. We also study purely deterministic rules, which are the $\alpha \to 0$ limit of the prior, where every row $f(\cdot \mid c)$ is a point mass.

*Example.* The classical *elementary cellular automata* (Wolfram, 1983) have $d = 1$, $V = 2$ and $k = 3$, with offsets $(-1, 0, 1)$. In $d = 2$ the *von Neumann* neighborhood has $k = 5$, the offsets $(0, 0)$, $(\pm 1, 0)$ and $(0, \pm 1)$, while the *Moore* neighborhood used in Conway's Game of Life (Conway, 1970) has $k = 9$, the offsets $(a, b)$ with $a, b \in \{-1, 0, 1\}$. A neighborhood of size $k$ over $V$ states admits $V^k$ configurations and therefore $V^{(V^k)}$ deterministic rules, which is 256 in the elementary case, $2^{32}$ for von Neumann and $2^{512}$ for Moore. Stochastic rules assign a distribution over $\mathcal{S}$, so they form a continuum of dimension $V^k(V - 1)$.

### 3.2 Sequence flattening and spatial routing

A trajectory is presented to the model as a single token stream. Within a time step the cells are emitted in row-major order, with the last coordinate varying fastest, and a `ROW_SEP` closes each innermost row, except that the last row of a step is closed by a `TIME_SEP` in place of its `ROW_SEP`. One step therefore reads $|\mathcal{G}|$ cell tokens and $|\mathcal{G}|/W_d$ separators:

$$\underbrace{x_{t,(0,0)}, \ldots, x_{t,(0,W_d-1)}}_{\text{row } 0}, \texttt{ROW_SEP}, \ \ldots, \ \underbrace{x_{t,(m,0)}, \ldots, x_{t,(m,W_d-1)}}_{\text{row } m=|\mathcal{G}|/W_d-1}, \texttt{TIME_SEP},$$

for $t = 0, \ldots, T - 1$ in order, so the stream has length $N = T(|\mathcal{G}| + |\mathcal{G}|/W_d)$. In $d = 1$ each time step is a single row, so the only separator is the `TIME_SEP`. The layout is an order-preserving injection $\rho : [T] \times \mathcal{G} \to [N]$, and we give the stream indexing in full in Appendix C. The flattening preserves adjacency along the innermost axis while obscuring adjacency along the others; the $k$ tokens in the parent configuration need not form a contiguous block.

In the transformer architecture, the cell token $(t, z)$ predicts the cell that follows it in the stream, one step along the row. We call such a token a *query cell* and the cell it predicts its *target cell*. We write $\hat{u}_d$ for this step and call it the *successor offset*; it is the $d$-th coordinate vector of the torus $\mathcal{G}$. A token therefore plays two roles. To make its *current prediction*, it must access the $k$ parents of its target cell $z + \hat{u}_d$, which lie at offsets $D_K + \hat{u}_d$ from $z$. To serve as a *precedent* for later predictions, it must also represent its own $k$ parents, which lie at offsets $D_K$. We denote the former set by $D_Q$.

**Definition 2** (Routing set)**.** For a query cell $(t, z)$, with target cell $z + \hat{u}_d$, the *query offsets* are $D_Q \coloneqq D_K + \hat{u}_d$, and the *routing set* is $\mathcal{U} \coloneqq D_K \cup D_Q = D_K \cup (D_K + \hat{u}_d) \subseteq \mathcal{G}$.

*Example.* On grids large enough to avoid wraparound, the elementary 1D neighborhood has $D_K = \{-1, 0, 1\}$ and $k = 3$, so $D_Q = \{0, 1, 2\}$ and $\mathcal{U} = \{-1, 0, 1, 2\}$, with $|\mathcal{U}| = 4$. The 2D von Neumann neighborhood has $k = 5$ and $|\mathcal{U}| = 8$; the 2D Moore neighborhood has $k = 9$ and $\mathcal{U} = \{-1, 0, 1\} \times \{-1, 0, 1, 2\}$, with $|\mathcal{U}| = 12$.

Generally, $k \leq |\mathcal{U}| \leq \min(2k, |\mathcal{G}|)$. Figure 2 shows both offset sets in a trained model. In these examples $D_K$ is symmetric but $\mathcal{U}$ is not, because $D_Q$ is shifted by $\hat{u}_d$. Along the innermost axis, $\mathcal{U}$ extends one step further in the positive direction than $D_K$, but is the same along other axes.

### 3.3 The model

We employ a two-layer causal decoder-only transformer (Vaswani et al., 2017) with learned absolute positional embeddings. Position $n$ of the token stream has residual-stream vector $x_n^{(0)} = W_E e_n + W_P p_n \in \mathbb{R}^{d_{\text{model}}}$, where $e_n$ is the one-hot of the token at position $n$ and $W_P$ embeds the learned absolute position $p_n \in \mathbb{R}^{d_p}$ (column $n$ of $P \in \mathbb{R}^{d_p \times N}$) into the residual stream. Attention weights

$$\alpha_{n,n'}^{(l)} \ = \ \text{softmax}_{n' \leq n}\Big(\big\langle W_Q^{(l)} x_n^{(l-1)}, \ W_K^{(l)} x_{n'}^{(l-1)} \big\rangle\Big).$$

at layer $l \in \{1, 2\}$ determine the output $\sum_{n' \leq n} \alpha_{n,n'}^{(l)} W_V^{(l)} x_{n'}^{(l-1)}$ which is projected back into the stream (optionally followed by a two-layer ReLU MLP of hidden width $d_{\text{MLP}}$). For simplicity of exposition, we absorb scaling factors into $W_Q$ and $W_K$.

Our analysis focuses on the learned positional table $P$, such as the one used in the original Vision Transformer (Dosovitskiy, 2020). Its columns are free parameters and supply no geometry of their own, so any spatial structure in a trained model must have been inferred from data. Appendix C.3 compares alternative positional embedding schemes.

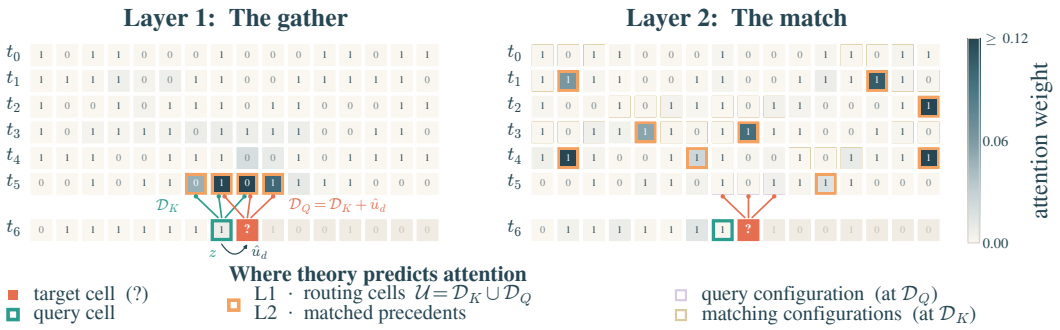


Figure 2: **Layer 1 gathers the routing cells; Layer 2 retrieves matching precedents.** Attention weights from the shared-head model on 1D $V{=}2, k{=}3$ (elementary cellular automata), Wolfram Rule 110. The query cell at $z$ predicts the target cell at $z{+}\hat{u}_d$ (marked "?"). Layer 1 (left) concentrates on the four routing cells in the previous timestep, at offsets $\mathcal{U} = D_K \cup D_Q$ relative to $z$, where $D_Q = D_K + \hat{u}_d$ gives the target cell's parent offsets. Layer 2 (right) concentrates on matching precedents: earlier cells whose parent configurations (at offsets $D_K$) match the query configuration (at offsets $D_Q$). Their states are used to predict the target cell.

### 3.4 Resolvable queries and the Bayes target

The kernel is drawn from the Dirichlet prior of Definition 1, so the best possible prediction at a token is based on the posterior, given everything earlier in the stream. It has a closed form: count the earlier occurrences of the query configuration, and how often each state followed them. Denote

$$n_c := \#\{(t', z') : t' \geq 1,\ \rho(t', z') \leq \rho(t, z),\ P_{t',z'} = c\}, \qquad n_{c,y} := \#\{\text{those with } x_{t',z'} = y\}.$$

**Theorem 3** (Bayes posterior predictive)**.** *For the Dirichlet prior and serialization (§3.2), the optimal predictive distribution at a token with query configuration $c$, given all earlier observations, is*

$$\widehat{P}_\alpha(y \mid c) \;=\; \frac{n_{c,y} + \alpha}{n_c + V\alpha}, \quad y \in \mathcal{S}.$$

The proof and further discussion are provided in Appendix D.

**Definition 4** (Resolvable query, coverage)**.** Configuration $c \in \mathcal{S}^k$ is *covered* at $(t, z)$ if $P_{t',z'} = c$ for some $(t', z')$ satisfying $\rho(t', z') \leq \rho(t, z)$, i.e. if $n_c > 0$. A query cell $(t, z)$ defines the *query configuration* $Q_{t,z} := P_{t,\, z+\hat{u}_d}$, the parent configuration of its target cell, which determines the target state distribution $f(\cdot \mid Q_{t,z})$. A query is *resolvable* if $Q_{t,z}$ is covered (Appendix D.1).

If a query configuration $c$ has not occurred earlier in the stream, then $n_c = 0$ and $\widehat{P}_\alpha(\cdot \mid c)$ reduces to the prior mean $1/V$, since there are no observations on $f(\cdot \mid c)$. In other words, the context only determines the target distribution if resolvability holds. In Section 4, we focus on resolvable queries.

## 4 Spatial Induction Heads: Construction and Optimality

### 4.1 The spatial induction head

For spatial data, the model must construct the relevant neighborhood at the query cell before matching. We call the resulting two-stage circuit a *spatial induction head*.

**Definition 5** (Spatial induction head)**.** A two-layer circuit is a *spatial induction head*, if at every resolvable query, it predicts in two stages:

(i) *The gather.* Layer 1 assembles the spatial neighborhood information needed at the query cell.
(ii) *The match.* Layer 2 compares the resulting configuration against earlier configurations in the context and averages over matching precedents to predict the next state.

In our cellular-automaton setting, the gather makes the states of all $|\mathcal{U}|$ routing cells, at offsets $\mathcal{U} = D_K \cup D_Q$, available at the query cell. These states serve two roles in Layer 2: the query reads

the offsets $D_Q$, which form the query configuration, the parent configuration of the target cell, while the key at an earlier token reads the offsets $D_K$, which form that token's own parent configuration. Layer 2 then performs the induction-style matching operation: the query configuration is compared against earlier parent configurations. Those that agree with it are the *matching configurations*, the cells carrying them are the *matching precedents*, and the states that followed those precedents are aggregated to form the prediction. Figure 2 shows the gather and match stages in a trained model.

### 4.2 Two constructions of the gather

A standard transformer can realize the gather in two ways: it can use one attention head per routed offset, so that attention itself keeps the gathered states separate, or it can gather all offsets with a single head and use an MLP to separate them afterward. Both constructions rely on positional addressing to locate the required cells using learned absolute positional embeddings.

**Lemma 6** (Addressing)**.** *For any nonempty target offset set $\mathcal{N} \subseteq \mathcal{G}$, a Layer-1 positional score which ranks the previous-step cells at offsets $\mathcal{N}$ from the query above all other visible cells can be realized with $d_p = d_S + 2$, where the spatial addressing width is $d_S = O(|\mathcal{N}|\, d \log |\mathcal{N}|)$ in general or $d_S = 2d$ when $|\mathcal{N}| = 1$. The positional dimension is therefore independent of $|\mathcal{G}|$ and $T$.*

The two gather constructions use this addressing primitive differently: Construction I applies it separately to each singleton offset $\{\delta\}$, while Construction II applies it once to the full routing set $\mathcal{U}$.

**Theorem 7** (Spatial gather constructions)**.** *The Layer-1 gather of $\mathcal{U}$ has two realizations, both at widths independent of $|\mathcal{G}|$ and $T$:*

- *(I)* ***Dedicated-heads construction.*** *$|\mathcal{U}|$ heads, one per offset, and no Layer-1 MLP, with $d_p = 2d+3$ and $d_{\text{model}} = (|\mathcal{U}| + 2)V + 2d + 6$. It recovers each routed state to any accuracy.*
- *(II)* ***Shared-head construction.*** *One head and an MLP which separates the routed states by offset, with $d_{\text{MLP}} \le 2(V-1)|\mathcal{U}|^3 + 3$ and $d_p \le O(|\mathcal{U}|\, d \log|\mathcal{U}|) + |\mathcal{U}|^3 + 3$. It recovers them exactly.*

A related head/nonlinearity tradeoff appears for sequential $k$-gram induction (Ekbote et al., 2026); here, the additional difficulty is preserving the identities of spatially separated offsets after routing.

**Proof sketch.** In Construction I, head $\delta$ spatially addresses the cell at offset $\delta \in \mathcal{U}$ and writes its state directly into a dedicated $V$-dimensional block of the residual stream. Since the heads write into disjoint blocks, the offsets are already separated after attention and no Layer-1 MLP is needed. Construction II uses nonlinear capacity instead of separate heads. A single head attends jointly to the cells at all offsets in $\mathcal{U}$. The cell state is encoded in the attention weight, while the value carries a positional label identifying the routing cell. A single global labeling guarantees these $|\mathcal{U}|$ routing cells remain mutually distinguishable at any query. Finally, the Layer-1 MLP uses the query's positional information to decode the aggregate into one state block per offset. The Fourier construction underlying the addressing bound, the positional labeling scheme for the shared-head construction, and the exact transformer parameter constructions are given in Appendices E, F, and G.

### 4.3 Matching and Bayes-optimal prediction

The gathered routing set feeds into a Layer-2 matching head. The query selects the gathered blocks indexed by $D_Q$, while each earlier key selects the blocks indexed by $D_K$. Their inner product, scaled by $\beta > 0$, measures agreement between the query configuration and an earlier parent configuration.

**Proposition 8** (Hard-attention matching as Bayesian counting)**.** *Suppose a single Layer-2 attention head assigns the same score $a$ to every precedent whose parent configuration matches the current query, and $-\infty$ to all other positions. With no Layer-2 MLP, the head outputs the empirical conditional transition law $(n_{c,y}/n_c)_{y\in\mathcal{S}}$. Augmenting the context with a single* prior token*, of value $\frac{1}{V}\mathbf{1}$ and score $a + \log(V\alpha)$, makes the output exactly $\widehat{P}_\alpha(y \mid c) = \frac{n_{c,y}+\alpha}{n_c + V\alpha}$, the posterior of Theorem 3.*

**Proof sketch.** For one-hot state blocks, the query–key inner product counts matching offsets, so an exact configuration match attains the maximal score. If all matches receive the same score and nonmatches are excluded, softmax is uniform over the $n_c$ matching precedents. Since each value carries the state that followed its configuration, their average is exactly the empirical conditional

transition law. The prior token contributes total unnormalized weight $V\alpha$ with uniform value $\frac{1}{V}\mathbf{1}$, adding $\alpha$ to every state count and recovering the posterior of Theorem 3. □

A finite softmax need not realize this ideal matching exactly: nonmatching positions retain some attention mass, and approximation error from Layer 1 can make the scores of matching positions slightly unequal. The end-to-end result below controls both effects.

**Theorem 9** (End-to-end spatial induction; formally Theorem H.9)**.** *Let the Layer-1 gather construction have per-block error $\bar{\epsilon}$, let $\beta$ denote the Layer-2 query–key scale, and write $\Delta := \beta\big(1-(2k+2)\bar{\epsilon}\big)$ for the resulting match–nonmatch separation. Suppose the Layer-2 head also discounts separators and initial-step cells. At every resolvable query with $n_c$ matching precedents:*

*(i)* Stochastic rules. *Fix $\kappa \in (0,1]$, assume $N \geq V\alpha$, and include the prior token of Proposition 8. If $\Delta \geq \log\big(4N/((n_c + V\alpha)\kappa)\big)$ and $\bar{\epsilon} \leq \kappa/(16k\Delta)$, then the Layer-2 attention output $p$, read directly as a distribution, satisfies $D_{\mathrm{KL}}\big(\widehat{P}_\alpha(\cdot \mid c) \,\big\|\, p\big) \leq \kappa$.*

*(ii)* Deterministic rules. *In the deterministic limit $\alpha \to 0$, if $\bar{\epsilon} < 1/(2k+2)$ and $\Delta > \log\big(2\sqrt{2}\,(N-n_c)/n_c\big)$, then the predicted next state is exactly correct.*

The two cases place different demands on the matching scores. For a deterministic rule, every matching precedent has the same next state, so a sufficient gap between matches and nonmatches guarantees the correct prediction. For a stochastic rule, the relative weights *among* matching precedents also matter: their uniform average is what recovers the transition counts. Layer-1 error perturbs these matched scores by at most $2\beta k\bar{\epsilon}$, so $\beta$ must be large enough to suppress nonmatches without amplifying this dispersion too much. The two conditions in Theorem 9 balance these effects.

# 5 Experiments

## 5.1 Experimental setup

We train two layer transformers, then evaluate them on rules held out from training to measure their ability for spatial in-context learning. We also examine whether the learned internal representations are consistent with the two-stage structure of the theoretical constructions. We use three settings: 1D $V{=}2, k{=}3$ (elementary cellular automata), 1D $V{=}3, k{=}3$, and 2D von Neumann (VN) $V{=}2, k{=}5$. The model receives only the first $T_0$ time steps of a trajectory generated under an unknown rule (not the rule table), and its weights are frozen at test time. Predictions are scored on cells from time step $t = T_0$ through $t = T{-}1$. Every evaluation trajectory is rejection-sampled so that all $V^k$ neighborhood configurations appear in the $T_0$ context rows, so every scored query is resolvable. Full details of data generation, training, and evaluation are in Appendix I.

**Rules.** (1) *Deterministic*: each trajectory draws one rule uniformly from the corresponding split's rule pool; training, validation, and test pools are mutually disjoint, so every val or test trajectory is generated by a rule the model has never seen during training. (2) *Stochastic*: each trajectory draws a fresh rule by independently sampling a Dirichlet$(1, \ldots, 1)$ transition row for every neighborhood configuration; because the Dirichlet is continuous, no two trajectories share the same rule.

Table 1: **Test-set prediction accuracy on held-out rules.** Deterministic metrics use the ground-truth history (teacher forcing, TF) or greedy autoregressive (AR) rollout. Stochastic AR cell accuracy is reported for the model and the $k$-gram sampling baseline. Stochastic AR KL is $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{model}})$ between the $k$-gram and model predictive distributions on the model's rollout.

| | | Deterministic | | Stochastic | |
|---|---|---|---|---|---|
| Task | Arch. | TF cell acc | AR cell acc | AR cell acc ($k$-gram) | AR KL |
| 1D $V{=}2,\ k{=}3$ | Shared-head | 1.0000 | 0.9999 | 0.5496 (0.5485) | 0.0019 |
| | Dedicated-heads | 1.0000 | 1.0000 | 0.5488 (0.5485) | 0.0044 |
| 1D $V{=}3,\ k{=}3$ | Shared-head | 1.0000 | 1.0000 | 0.3476 (0.3475) | 0.0026 |
| | Dedicated-heads | 1.0000 | 1.0000 | 0.3475 (0.3475) | 0.0004 |
| 2D VN $V{=}2,\ k{=}5$ | Shared-head | 1.0000 | 1.0000 | 0.5201 (0.5197) | 0.0013 |
| | Dedicated-heads | 1.0000 | 0.9999 | 0.5191 (0.5197) | 0.0038 |

**Architectures.** We instantiate the two gather constructions of Theorem 7: the *Dedicated-heads* construction ($|\mathcal{U}|$ L1 attention heads with no L1 MLP) and the *Shared-head* construction (one L1 attention head with L1 MLP). Both share a single L2 attention head and a two-layer L2 MLP.

**Metrics.** For deterministic rules, we report *TF cell accuracy* (each predicted cell conditioned on the ground-truth history) and *AR cell accuracy* (greedy autoregressive rollout scored against the ground-truth trajectory). For stochastic rules, *AR cell accuracy* is the fraction of cells in a temperature-1 rollout that match the ground truth. The ceiling is well below 1; we also report a posterior-predictive sampling baseline, an independent rollout which samples each cell from the add-one-smoothed $k$-gram posterior predictive (its mode maximizes expected one-step accuracy). We additionally report *AR* $D_{\mathrm{KL}}(q_{kgram} \| q_{model})$, computed at each cell during autoregressive generation between the model's predictive distribution and the causal $k$-gram distribution on the model's own rollout.

### 5.2 Results

**Can the model recover unseen rules?** Table 1 reports test-set performance across three settings and both architectures. On deterministic rules, all models reach near-perfect cell accuracy under both teacher forcing and autoregressive rollout. On stochastic rules, both architectures match the $k$-gram baseline in AR cell accuracy, and all AR KL values fall below $0.005$ nats. Both architectures thus perform accurate in-context prediction for previously unseen rules across all three settings.

**Do attention heads attend where each construction predicts?** We next ask whether the model recovers the rule using the constructed two stage gather-and-match. We first examine the attention mass. Figure 3 (left) reports the mean attention mass on the construction-predicted positions. Both architectures place substantial mass on these positions across all three settings. To further test whether internal model computations rely on these positions, we examine what happens when we mask attention to these cells. For both Shared-head and Dedicated-heads models, we mask L1 attention to routing cells or L2 attention to matching precedents at evaluation time, using an equal number of randomly selected cells as controls (Figure 3, right). Across all three settings, masking attention to the theory-predicted cells reduces cell accuracy, while the random controls have little effect. This supports the role of both the routing and matching stages in the trained models' predictions.

**Do trained models exhibit the predicted layerwise structure?** We train linear probes on frozen post-residual representations after embedding, L1 attention, L1 MLP, L2 attention, and L2 MLP. For two targets: (1) the *neighborhood configuration*, the $k$-cell neighborhood that determines the next cell value; (2) the *$k$-gram predictive distribution*, the conditional distribution over the next cell value computed from causal $k$-gram counts in the trajectory prefix with add-one smoothing (the probe minimizes $D_{\mathrm{KL}}(q_{\mathrm{kgram}} \| q_{\mathrm{probe}})$). Probe training, validation, and test rules are disjoint. The constructions make the configuration linearly readable after the L1 MLP in Shared-head models and after L1 attention in Dedicated-heads models, and the predictive distribution after L2 attention in both via the simplex readout. Our logistic probes are an empirical diagnostic of this structure, and both architectures match it across all three settings (Figure 4). Figure 5 visualizes the layerwise structure

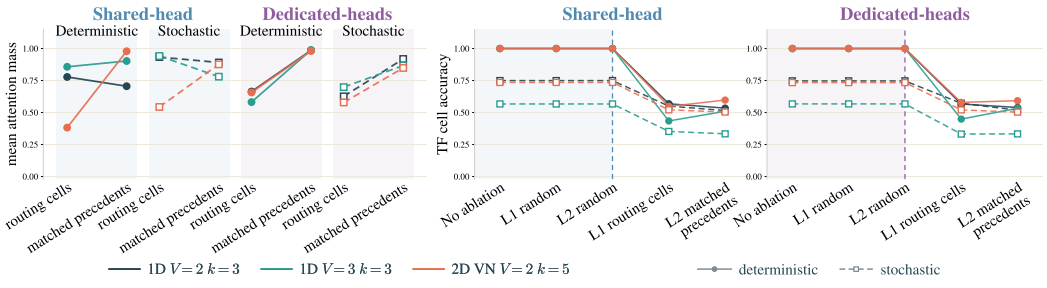


Figure 3: **Attention alignment and ablation.** *Left:* Mean attention mass on the positions predicted by each construction: L1 routing cells at offsets $\mathcal{U} = D_K \cup D_Q$, and L2 matching precedents, earlier cells whose parent configurations match the query configuration. *Right:* Cell accuracy when attention to the L1 routing cells or L2 matching precedents is masked at evaluation time, compared with no ablation and controls that mask attention to the same number of randomly selected cells.

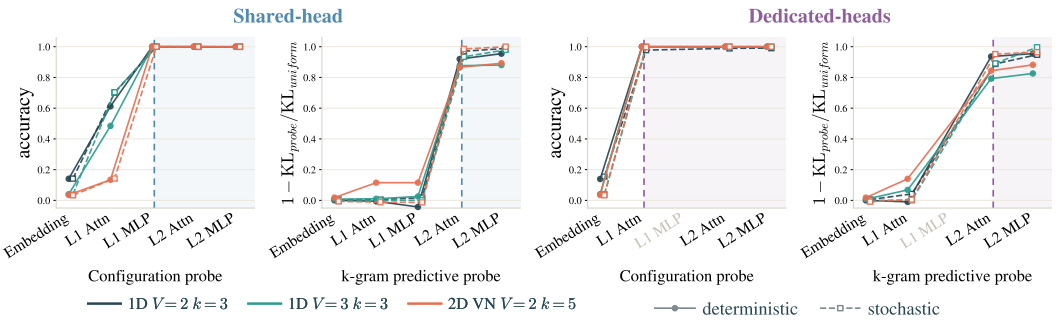


Figure 4: **Linear probe results.** Higher is better. The configuration probe measures classification accuracy for the $k$-cell neighborhood; the k-gram predictive probe measures $1 - \overline{D_{\mathrm{KL}}(q_{\mathrm{kgram}} \| q_{\mathrm{probe}})} \,/\, D_{\mathrm{KL}}(q_{\mathrm{kgram}} \| \mathrm{Uniform})$, which the axis label abbreviates as $1 - \mathrm{KL}_{\mathrm{probe}}/\mathrm{KL}_{\mathrm{uniform}}$. Solid lines: deterministic; dashed: stochastic. Vertical lines mark the stage at which the theoretical construction places each target.

identified by probes, using t-SNE on the Shared-head model on deterministic 1D $V=2, k=3$. After L1 MLP, representations cluster by neighborhood configuration; after L2 attention, by output value.

## 6 Conclusion

We studied how transformers perform in-context learning when the dependencies required for prediction are spatial. We introduced the *spatial induction head*, a two-layer gather-and-match circuit in which the first layer assembles the relevant spatial context from a serialized sequence and the second matches that context against precedent to predict the next state. We gave two explicit realizations of the gather and showed that the positional width required for spatial routing depends only on the local neighborhood and spatial dimension, not on the total grid volume or trajectory horizon. We further showed that the matching computation realizes Bayesian counting: the end-to-end circuit can approach the Bayes posterior predictive arbitrarily closely for stochastic rules and predicts exactly for deterministic rules. Experiments on 1D and 2D cellular automata demonstrate generalization to held-out rules and provide mechanistic evidence consistent with the gather-and-match computation.

Our analysis focuses on cellular automata, whose finite-state local dynamics provide a particularly tractable setting for isolating spatial induction. More broadly, by separating spatial routing from precedent matching, our construction provides a framework for understanding induction mechanisms in more general types of data with spatial structure.

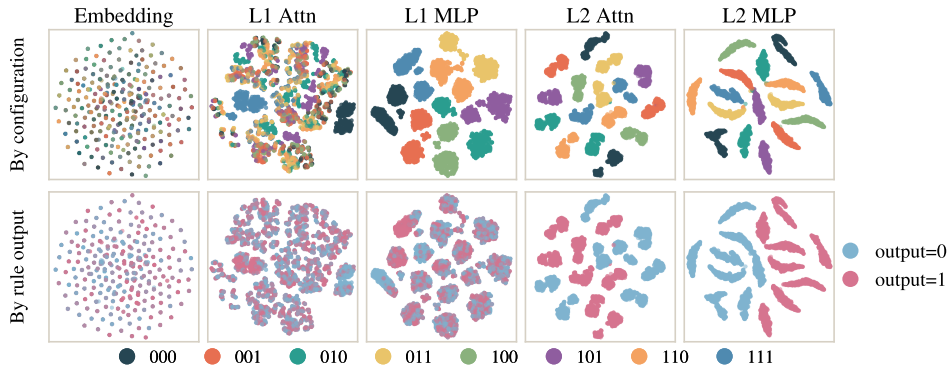


Figure 5: **t-SNE visualization of frozen representations** from the Shared-head model on deterministic 1D $V=2, k=3$. Top row (colored by neighborhood configuration): configurations cluster after the L1 MLP. Bottom row (colored by output value): outputs cluster after L2 attention.

## AI USE STATEMENT

We used generative AI tools to assist with implementing and debugging experimental code, excluding training code. The authors specified the intended methods, reviewed and revised the AI-assisted code, and checked its behavior through experiments.

We also used generative AI tools to help draft illustrative numerical examples, explanatory remarks, and reading guidance in the appendices, and to edit the manuscript for clarity and readability. The authors reviewed and revised all AI-assisted text for scientific accuracy, appropriate attribution, and correct formatting.

The authors take responsibility for the final content of this work, including its text, claims, code, and reported results.

## REPRODUCIBILITY STATEMENT

Appendix I provides details supporting reproduction of our experiments: data generation and rule splits in Appendix I.1, model architectures and training settings in Appendix I.2, and evaluation protocols and metric definitions in Appendix I.3. It also reports the full predictive results and describes the attention analyses, ablations, layerwise probes, and representation visualizations.

## REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Roman Bachmann, Jesse Allardice, David Mizrahi, Enrico Fini, Oğuzhan Fatih Kar, Elmira Amirloo, Alaaeldin El-Nouby, Amir Zamir, and Afshin Dehghan. Flextok: Resampling images into 1d token sequences of flexible length. In *Forty-second International Conference on Machine Learning*, 2025.

Jaime A Berkovich, Noah S David, and Markus J Buehler. Automatagpt: Forecasting and ruleset inference for two-dimensional cellular automata. *arXiv preprint arXiv:2506.17333*, 2025.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

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.

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.

Mikhail Burtsev. Learning elementary cellular automata with transformers. *arXiv preprint arXiv:2412.01417*, 2024.

Frank Cole, Yulong Lu, Tianhao Zhang, and Yuxuan Zhao. In-context learning of linear dynamical systems with transformers: Error bounds and depth-separation. *arXiv preprint arXiv:2502.08136*, 3(7), 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

John Conway. Conway's game of life. *Scientific American*, 223:120–123, 1970.

Yijia Dai, Zhaolin Gao, Yahya Sattar, Sarah Dean, and Jennifer Sun. Pre-trained large language models learn to predict hidden markov models in-context. *Advances in Neural Information Processing Systems*, 38:28877–28923, 2026.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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.

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.

Roberto Fernández, Pierre-Yves Louis, and Francesca R Nardi. Overview: Pca models and issues. *Probabilistic Cellular Automata: theory, applications and future perspectives*, pp. 1–30, 2018.

Angeliki Giannou, Liu Yang, Tianhao Wang, Dimitris Papailiopoulos, and Jason D Lee. How well can transformers emulate in-context newton's method? *arXiv preprint arXiv:2403.03183*, 2024.

András Hajnal. Proof of a conjecture of p. erdős. *Combin. Theory Appl.*, 2:601, 1970.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Chris G Langton. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: nonlinear phenomena*, 42(1-3):12–37, 1990.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

Toni JB Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher Earls. Llms learn governing principles of dynamical systems, revealing an in-context neural scaling law. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 15097–15117, 2024.

Ashok V Makkuva, Marco Bondaschi, Chanakya Ekbote, Adway Girish, Alliot Nagle, Hyeji Kim, and Michael Gastpar. Local to global: Learning dynamics and effect of initialization for transformers. *Advances in Neural Information Processing Systems*, 37:86243–86308, 2024a.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*, 2024b.

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

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

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.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Nived Rajaraman, Marco Bondaschi, Kannan Ramchandran, Michael Gastpar, and Ashok V Makkuva. Transformers on markov data: Constant depth suffices. *Advances in Neural Information Processing Systems*, 37:137521–137556, 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.

Audrey Terras. *Fourier analysis on finite groups and applications*. Number 43. Cambridge University Press, 1999.

Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. *Advances in Neural Information Processing Systems*, 37:26941–26975, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Stephen Wolfram. Statistical mechanics of cellular automata. *Rev. Mod. Phys.*, 55:601–644, Jul 1983. doi: 10.1103/RevModPhys.55.601. URL https://link.aps.org/doi/10.1103/RevModPhys.55.601.

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.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18431–18441. IEEE, 2025.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

## APPENDIX CONTENTS

**A Roadmap and reading paths** **14**

**B Notation and conventions** **15**

B.1 Conventions . . . . . 15

B.2 Symbols . . . . . 15

B.3 Overloaded notation . . . . . 17

B.4 The experimental settings . . . . . 18

**Part I — The task**

**C Serialization and limits of sequence-relative routing** **19**

C.1 Formal serialization and indexing . . . . . 19

C.2 Boundary and separator queries . . . . . 20

C.3 Limits of displacement-only positional biases . . . . . 21

**D Coverage and Bayes-optimal prediction** **24**

D.1 Coverage and resolvable queries . . . . . 24

D.2 Bayes-optimal posterior prediction . . . . . 26

**Part II — Spatial routing primitives**

**E Spatial addressing with grid-independent width** **27**

E.1 Fourier representation of spatial attention scores . . . . . 27

E.2 Isolating a set of offsets . . . . . 29

E.3 Attention concentration from a score gap . . . . . 32

E.4 Selecting spatial offsets in the preceding time step . . . . . 32

E.5 Example: addressing the 2D von Neumann routing set . . . . . 35

**F Colorings for Shared-head decoding** **36**

F.1 Admissible colorings and the Cayley graph . . . . . 36

F.2 Bounds on the number of colors . . . . . 37

F.3 Colorings for the experimental grids . . . . . 37

F.4 Coloring interface for Shared-head decoding . . . . . 38

**Part III — Layer 1: the spatial gather**

**G Layer 1: spatial gather constructions** **39**

G.1 Target representation and residual-stream layout . . . . . 40

G.2 Separator gating and finite-softmax error . . . . . 42

G.3 Construction I: One head per offset . . . . . 43

G.4 Construction II: A shared head with MLP decoding . . . . . 44

G.5 The spatial gather interface . . . . . 48

G.6 Construction sizes for the experimental settings . . . . . 49

**Part IV — Layer 2 and end-to-end guarantees**

**H Layer 2: Bayesian matching and end-to-end guarantees** **50**

H.1 The configuration-matching head . . . . . 50

H.2 Exact counting and Dirichlet smoothing . . . . . 51

H.3 Separation of matches from non-matches . . . . . 52

H.4 Matched-score dispersion and predictive error . . . . . 54

H.5 The key-validity gate . . . . . 55

H.6 End-to-end spatial induction . . . . . 56
H.7 Choosing finite parameters . . . . . 58

**Part V — Experiments**

**I Experimental methods and additional results** **59**

I.1 Data generation and rule splits . . . . . 59
I.2 Models and training . . . . . 60
I.3 Evaluation protocol and metrics . . . . . 61
I.4 Full predictive results . . . . . 62
I.5 Attention-alignment analysis . . . . . 62
I.6 Attention ablation . . . . . 63
I.7 Layerwise probe protocol . . . . . 63
I.8 Representation visualizations with t-SNE . . . . . 64
I.9 Qualitative rollout setup for Figure 1 . . . . . 64

## A Roadmap and reading paths

The appendices are organized thematically. Part I formalizes the task, detailing grid serialization, prediction correctness, and the limitations of standard positional schemes. Part II establishes grid-based mathematical foundations, yielding two interface corollaries. Parts III and IV describe Layers 1 and 2 of the architecture, respectively. Part V outlines experimental methods.

- **Notation** (Appendix B). Symbol reference and running example.
- **Serialization and limits of sequence-relative routing** (Appendix C), *Part I*. Formalizes flattening (Section 3.2) and separator functions. Proves that sequence-relative biases (Definition C.5) cannot route local neighborhoods on a torus in a single layer, motivating the spatial addressing in Part II.
- **Coverage and Bayes-optimal prediction** (Appendix D), *Part I*. Defines resolvable queries and the optimal predictor for empirical accuracy.
- **Spatial addressing with grid-independent width** (Appendix E), *Part II*. Demonstrates that $O(|\mathcal{N}|d\log|\mathcal{N}|)$ coordinates suffice to score routing cells at offsets $\mathcal{U}$ above other visible keys, independent of grid size.
- **Colorings for Shared-head decoding** (Appendix F), *Part II*. Formulates an injective grid labeling scheme for translates $z+\mathcal{U}$, corresponding to a Cayley graph proper coloring.
- **Layer 1: spatial gather constructions** (Appendix G), *Part III*. Details the forward pass, residual stream, and weights. Gathers are performed using either $|\mathcal{U}|$ Dedicated-heads or a single Shared-head with a separating MLP.
- **Layer 2: Bayesian matching and end-to-end guarantees** (Appendix H), *Part IV*. Constructs the matching mechanism. Shows that softmax over an equally scored set implements a counter, yielding the Bayes predictive distribution. Bounds the required separation $\Delta$ against Layer 1's dispersion error $\eta$.
- **Experimental methods and additional results** (Appendix I), *Part V*. Details data generation, training, and evaluation protocols, including complete metric tables and mechanistic analyses.

**Reading paths.** The shortest sequence comprising the main result involves six statements. Read the Appendix G introduction, followed by the two gather theorems (Theorems G.9 and G.10) and their Layer 2 interface (Corollary G.12). Continue to the Appendix H introduction, the counting identity (Theorem H.2), and the end-to-end guarantee (Theorem H.9). Two prerequisites are referenced on demand: Corollary E.6 (positional addressing) and Corollary F.5 (colorings for Theorem G.10).

Table 2: Dependencies between construction stages.

| Stage result | Downstream usage |
|---|---|
| Corollary E.6 | Positional features for offset selection in previous time steps |
| Corollary F.5 | Distinct color slots and storage costs (Construction II) |
| Corollary G.12 | State blocks (approximate or exact) per routed offset |
| Theorem H.2 | Equal-weight matching for empirical counts and Dirichlet smoothing |
| Lemmas H.3, H.5, H.6 | Bounds on separation, dispersion, and predictive cost |
| Theorem H.9 | Finite-parameter prediction guarantee |

Suggested reading paths by topic:

- *Circuit construction:* Appendices G and H (Parts III and IV).
- *Mathematical foundations:* Part II.
- *Formal problem setting:* Part I.
- *Symbol reference:* Appendix B.

# B Notation and conventions

This section catalogs symbols defined throughout the text; Table 3 provides their definitions and points of introduction. Section B.3 clarifies overloaded symbols, and Section B.4 instantiates task parameters for the three experimental settings.

## B.1 Conventions

Coordinates, states, time steps, sequence positions, and residual blocks ($B_0, \ldots, B_4$) are 0-indexed; dimensions, offsets, and heads are 1-indexed. Let $[n] \coloneqq \{0, \ldots, n-1\}$. Grid coordinate arithmetic is modulo the respective side length, making $\mathcal{G}$ a torus such that $z + \delta$ always identifies a valid cell.

Grid side lengths are denoted $W_1, \ldots, W_d$, and its *volume* is $|\mathcal{G}| = \prod_j W_j$.

A *row* designates an innermost-axis grid row, comprising the $W_d$ cells followed by a `ROW_SEP`. A *time step* $t$ corresponds to one automaton layer; a query at step $t$ accesses step $t-1$. The term *row* strictly retains its spatial meaning.

Vectors are column vectors. The notation $|\cdot|$ denotes set cardinality or the distance to 0 for group elements. The one-hot vector for coordinate $v$ is $e_v$; for a state $x$, this is written $e(x)$ to distinguish it from the dual-group generators $e_1, \ldots, e_d$.

## B.2 Symbols

Table 3: Symbols, and where each is fixed.

| symbol | meaning | fixed at |
|---|---|---|
| *Grid and dual group* | | |
| $d$ | number of spatial dimensions | Def. 1 |
| $W_j$ | side length of axis $j$ | Def. 1 |
| $W_{\max}$ | largest side length, $\max_j W_j$ | Lem. E.3 |
| $\mathcal{G}$, $\lvert\mathcal{G}\rvert$ | the torus $[W_1] \times \cdots \times [W_d]$; its volume $\prod_j W_j$ | Def. 1 |
| $z$, $z_j$, $\delta$ | a cell of $\mathcal{G}$; its $j$-th coordinate; an offset | Def. 1 |
| $\widehat{\mathcal{G}}$ | dual group, the characters of $\mathcal{G}$ | App. E.1 |
| $\omega$, $e_j$ | a character; the $j$-th generator of $\widehat{\mathcal{G}}$ | App. E.1 |
| $\varphi_\omega$, $\chi_\omega$ | phase $2\pi \sum_j \omega_j \delta_j / W_j$; the character $e^{i\varphi_\omega}$ | App. E.1 |

*continued on the next page*

*Table 3, continued*

| symbol | meaning | fixed at |
|---|---|---|
| $\operatorname{supp} \hat{g}$ | spectrum of $g$, the $\omega$ with $\hat{g}(\omega) \neq 0$ | App. E.1 |
| *States, rules and the prior* | | |
| $\mathcal{S}$, $V$ | state set $[V]$, and its size | Def. 1 |
| $T$, $t$ | horizon; a time step | Def. 1 |
| $T_0$ | context length, the number of time steps shown before scoring begins | §5 |
| $(t, z)$ | the cell $z$ at step $t$ | Def. 1 |
| $x_{t,z}$ | state of cell $z$ at step $t$ | Def. 1 |
| $f$ | the local kernel $\mathcal{S}^k \to \Delta^{V-1}$ | Def. 1 |
| $\alpha$ | Dirichlet concentration of the prior on $f$ | Def. 1 |
| $P_{t,z}$ | parent configuration of $(t, z)$ | Def. 1 |
| $Q_{t,z}$ | query configuration, $P_{t,z+\hat{u}_d}$ | Def. 4 |
| $c$ | a *matching configuration*: an earlier $P_{t',z'}$ equal to $Q_{t,z}$; its cell is a *matching precedent* | Def. 4 |
| $\widehat{P}_\alpha$ | Bayes posterior predictive | Thm. 3 |
| *Neighborhood and routing* | | |
| $(\delta_1, \ldots, \delta_k)$, $k$ | the neighborhood, a list of distinct offsets, and its size | Def. 1 |
| $D_K$, $D_Q$ | key offsets; query offsets $D_K + \hat{u}_d$ | Def. 1, 2 |
| $\hat{u}_d$ | successor offset, the $d$-th coordinate vector | Def. C.1 |
| $\mathcal{U}$ | routing set $D_K \cup D_Q$; the cells at these offsets are the *routing cells* | Def. 2 |
| $\mathcal{N}$ | a generic set of offsets to be isolated | Thm. E.4 |
| *Serialization* | | |
| $\rho_{\mathcal{G}}$, $r(z)$ | row-major index of $z$; its row | Def. C.1 |
| $\rho(t, z)$ | sequence position of the cell $(t, z)$ | Def. C.1 |
| $N$ | stream length $T(\lvert\mathcal{G}\rvert + \lvert\mathcal{G}\rvert/W_d)$ | Def. C.1 |
| $n$, $n'$ | a query position; a key position | §3.3 |
| $\tau(n)$, $\kappa(n)$ | virtual time step and virtual cell of position $n$ | App. C.1 |
| query cell | a cell token which poses a query | Def. 2 |
| target cell | the cell $\kappa(n) + \hat{u}_d$ which $n$ predicts | Lem. C.2 |
| `ROW_SEP` | separator closing an innermost row | Def. C.1 |
| `TIME_SEP` | separator closing a time step | Def. C.1 |
| *The model* | | |
| $W_E$, $e_n$, $P$, $p_n$ | token embedding; one-hot of the token at $n$; positional table; its column $n$ | §3.3 |
| $W_Q^{(l)}, W_K^{(l)}, W_V^{(l)}$ | query, key and value projections at layer $l$ | §3.3 |
| $g$, $\gamma$ | the positional kernel on $\mathcal{G}$; its isolation margin | Def. E.2 |
| $\Lambda$ | Layer-1 logit scale | Cor. E.5 |
| $\Gamma_1$ | Layer-1 separator gate strength | Prop. G.6 |
| $\Gamma_2$ | Layer-2 key-validity gate strength | Prop. H.7 |
| $\nu_n$ | key-validity indicator $s_n + \zeta_n$ | Def. H.1 |
| $\beta$ | Layer-2 query scaling | Def. H.1 |
| *Widths* | | |
| $d_{\text{model}}$ | residual-stream width | §3.3 |
| $d_p$ | width of the positional table, $d_S + 3$ for Construction I | Def. G.2 |
| $d_S$ | addressing dimension, $\lvert\operatorname{supp} \hat{g} \setminus \{0\}\rvert$ | §E.1 |
| $d_{\text{MLP}}$ | hidden width of the Layer-1 MLP | §3.3 |
| *Residual blocks and projections* | | |
| $B_0$ | separator indicator, auxiliary-token marker, and the constant 1 | Def. G.2 |

*continued on the next page*

*Table 3, continued*

| symbol | meaning | fixed at |
|---|---|---|
| $\mu_n$ | auxiliary-token marker $\mathbf{1}[n = \star]$, a coordinate of $B_0$ | Def. G.2 |
| $\sigma_\star$ | Layer-2 score of the auxiliary token $\star$ | Def. H.1 |
| $B_1$ | own state one-hot $e(x_{t,z})$ | Def. G.2 |
| $B_2$ | positional table entry $p_{t,z}$ | Def. G.2 |
| $\zeta_n$ | initial-step indicator $\mathbf{1}[\tau(n) = 0]$, a coordinate of $B_2$ | Def. G.2 |
| $B_2^{\mathrm{col}}$ | own-color block of $B_2$, zero at separators | Def. G.2 |
| $B_3$ | pattern scratch, $B_3^{\mathrm{pat}} \oplus B_3^{\mathrm{agg}}$ | Def. G.2 |
| $B_4$ | output slot, written by Layer 2 | Def. G.2 |
| $\Pi_{B_r}$ | projection onto block $B_r$ | Def. G.2 |
| $\iota_\delta$ | embedding of $\mathbb{R}^V$ into block $\delta$ of $B_3^{\mathrm{pat}}$ | Thm. G.9 |
| $\Sigma_Q, \Sigma_K$ | Layer-2 selection projections onto $\mathbb{R}^{kV}$ | Def. H.1 |
| *Coloring* | | |
| $c$ | a coloring $\mathcal{G} \to [M]$ | Def. F.1 |
| $M$ | number of colors used | Def. F.1 |
| $M^\star$ | chromatic number of $\Gamma_{\mathcal{U}}$ | Def. F.1 |
| $\Gamma_{\mathcal{U}}$ | Cayley graph on $(\mathcal{U} - \mathcal{U}) \setminus \{0\}$ | Def. F.1 |
| $\chi, K, M_K$ | periodic coloring; its subgroup; its color count $\lvert\mathcal{G}/K\rvert$ | Def. F.1, Lem. F.2 |
| $n_{\max}$ | most visible cells sharing one color | Thm. G.10 |
| *Budgets, gaps and counts* | | |
| $\bar{\varepsilon}$ | per-block error, the routing budget | Def. G.1; discharged by Lem. G.8 |
| $\varepsilon_{\mathrm{ret}}$ | retrieval error at score gap $\Delta$ | Cor. H.4 |
| $\Delta$ | Layer-2 separation $\beta(1 - (2k+2)\bar{\varepsilon})$, the gap between matching and non-matching keys | Lem. H.3 |
| $\eta$ | dispersion within the matched set, $2\beta k\bar{\varepsilon}$ | Lem. H.5 |
| $\kappa$ | divergence budget of the main theorem | Thm. H.9 |
| $D_{\mathrm{KL}}$ | Kullback–Leibler divergence | Thm. 9 |
| $n_c, n_{c,y}$ | times $c$ has been seen; times it was followed by $y$ | Def. 4 |
| $\mathcal{M}_{t,z}$ | the matched set | §H.1 |

### B.3 Overloaded notation

This work integrates harmonic analysis on finite abelian groups, graph coloring, and transformer architectures. To adhere to established conventions in these fields, eight symbols are overloaded. The surrounding context unambiguously determines the active definition, and no single equation mixes distinct definitions of the same symbol.

Table 4: The three experimental settings as tasks.

| | 1D, $V{=}2$ | 1D, $V{=}3$ | 2D von Neumann |
|---|---|---|---|
| $\mathcal{G}$ | $\mathbb{Z}_{16}$ | $\mathbb{Z}_{32}$ | $\mathbb{Z}_6^2$ |
| $d$, $W_{\max}$, $\lvert\mathcal{G}\rvert$ | 1, 16, 16 | 1, 32, 32 | 2, 6, 36 |
| $V$, $T$ | 2, 10 | 3, 12 | 2, 12 |
| $D_K$, $k$ | $\{-1, 0, 1\}$, 3 | $\{-1, 0, 1\}$, 3 | von Neumann, 5 |
| $\hat{u}_d$, $\lvert\mathcal{U}\rvert$ | 1, 4 | 1, 4 | $(0, 1)$, 8 |
| tokens per step, $N$ | 17, 170 | 33, 396 | 42, 504 |

| symbol | readings |
|---|---|
| $c$ | a *configuration* in $\mathcal{S}^k$ (Def. 4); a *coloring* $\mathcal{G} \to [M]$ (Def. F.1). |
| $\kappa$ | the *virtual cell index* $\kappa(n)$, always applied to a position (App. C.1); the *divergence budget* $\kappa \in (0, 1]$, always bare (Thm. H.9). |
| $B$ | a *residual block* $B_0, \ldots, B_4$, always subscripted (Def. G.2); the *bounding box* of a target set (Lem. E.3); the MLP *gate constant* $B_\delta = G_\delta$ of (40), subscripted by an offset. |
| $P$ | the *positional table* (§3.3); the *parent configuration* $P_{t,z}$ (Def. 1); the *axis factor* $P_j$ of a box kernel (Lem. E.3). |
| $Q$ | the *bilinear form* realizing a kernel (Lem. E.1); the *box kernel* $Q = \sum_j P_j$ (Lem. E.3); the *query configuration* $Q_{t,z}$ (Def. 4). |
| $M$ | the *number of colors* of an admissible coloring (Def. F.1), subscripted $M_K$ for a periodic one and starred $M^\star$ for the chromatic number; the *score matrix* in the proof of Lem. E.1, in an appendix which never colors anything; the *matched set* $\mathcal{M}_{t,z}$, calligraphic and subscripted (§H.1). |
| $\Delta$ | the *logit separation* (Lem. H.3); the *simplex* $\Delta^{V-1}$, always superscripted (Def. 1); the *maximum degree* in the Hajnal–Szemerédi bound (proof of Lem. F.3). |
| $G$ | the *spatiotemporal kernel* $G(\delta, s)$, always applied to an offset and a step displacement (proof of Cor. E.6, Stage 6); the grid $\mathcal{G}$ and the Cayley graph $\Gamma_{\mathcal{U}}$, both distinguished by font; the MLP *gain* $G_\delta$, always subscripted (proof of Thm. G.10). |

Additionally, $\alpha$ denotes the Dirichlet concentration (Definition 1), except when written as $\alpha^{(l)}_{n,n'}$, where it represents the attention weight (§3.3) with a layer superscript and positional subscripts.

### B.4 The experimental settings

The experiments (Appendix I) cover three neighborhood types: binary 1D ($V = 2, k = 3$), ternary 1D ($V = 3, k = 3$), and binary 2D von Neumann ($V = 2, k = 5$). Table 4 specifies the deterministic instances considered below, on $\mathbb{Z}_{16}$, $\mathbb{Z}_{32}$, and $\mathbb{Z}_6^2$, respectively. The stochastic experiments use the same neighborhood types but have different task parameters where indicated in Table 10; in particular, the stochastic ternary 1D setting uses $\mathbb{Z}_{16}$. Table 6 gives the colorings for the instances in Table 4 alongside the bounds from Appendix F. Table 8 (Appendix G) details the widths and scales required by the two constructions.

## Part I — The task

# C SERIALIZATION AND LIMITS OF SEQUENCE-RELATIVE ROUTING

This section comprises three parts. Section C.1 specifies the mapping from cells to sequence positions, extending it to separators via a virtual cell index. Section C.2 addresses boundary conditions introduced by row separators. Section C.3 proves that sequence-relative positional biases cannot strictly isolate required routing cells. Appendix E constructs an absolute positional table for spatial addressing.

## C.1 FORMAL SERIALIZATION AND INDEXING

**Definition C.1** (Row-terminated serialization)**.** Cells are emitted in row-major order:

$$\rho_{\mathcal{G}}(z) \;\coloneqq\; \sum_{j=1}^{d} z_j \prod_{l>j} W_l \;\in\; [|\mathcal{G}|], \qquad r(z) \;\coloneqq\; \big\lfloor \rho_{\mathcal{G}}(z)/W_d \big\rfloor \;\in\; [|\mathcal{G}|/W_d].$$

Here, $\rho_{\mathcal{G}}(z)$ is the flattened grid index of cell $z$, and $r(z)$ is the global index of the innermost row containing $z$, such that $\rho_{\mathcal{G}}(z) = r(z)W_d + z_d$. Let

$$\hat{u}_d \;\coloneqq\; (0, \ldots, 0, 1) \in \mathcal{G}$$

denote the *successor offset*, which corresponds to advancing one position within a row.

Every innermost row concludes with a separator: a `TIME_SEP` for the final row of a time step, and a `ROW_SEP` otherwise. Letting $(r, c)$ denote the cell in row $r$ with last coordinate $c$, one time step is serialized as:

$$\underbrace{\underbrace{x_{t,(0,0)}, \ldots, x_{t,(0,W_d-1)}}_{\text{row } 0}, \texttt{ROW_SEP}, \ldots, \underbrace{x_{t,(m,0)}, \ldots, x_{t,(m,W_d-1)}}_{\text{row } m=|\mathcal{G}|/W_d-1}, \texttt{TIME_SEP}}_{|\mathcal{G}|+|\mathcal{G}|/W_d \text{ tokens}}. \tag{1}$$

The complete stream comprises (1) for $t = 0, \ldots, T-1$, with total length

$$N \;\coloneqq\; T(|\mathcal{G}| + |\mathcal{G}|/W_d).$$

The cell $(t, z)$ occupies sequence position

$$\rho(t, z) \;=\; t(|\mathcal{G}| + |\mathcal{G}|/W_d) + \rho_{\mathcal{G}}(z) + r(z), \tag{2}$$

where the terms represent tokens in prior time steps, preceding cells in the current step, and preceding row separators. The map $\rho : [T] \times \mathcal{G} \to [N]$ is thus an order-preserving injection.

*Remark* (Final separator). By terminating the final time step with a `TIME_SEP`, every time step maintains equal length, preventing terminal case distinctions in (2).

*Example* (One dimension). In $d = 1$, each time step is one row, so $r(z) = 0$ and $|\mathcal{G}|/W_d = 1$. Writing $W = W_1$, one time step is $W$ cell tokens followed by one `TIME_SEP`, and

$$\hat{u}_d = 1, \qquad \rho(t, z) = (W + 1)t + z.$$

*Example* (Two dimensions). In $d = 2$, with grid dimensions $W_1 \times W_2$, the innermost row length is $W_2$. For a cell at $z = (z_1, z_2)$:

$$\rho_{\mathcal{G}}(z) = z_1 W_2 + z_2, \qquad r(z) = z_1.$$

Here, $r(z)$ extracts the row coordinate $z_1$. One time step contains $W_1$ rows of length $W_2$, requiring $W_1$ separators. The sequence position is therefore:

$$\rho(t, z) = t(W_1 W_2 + W_1) + (z_1 W_2 + z_2) + z_1.$$

**Causal masking.** The model employs a standard causal mask; a query at sequence position $n$ attends only to positions $n' \le n$. For cell tokens, $(t', z')$ is visible to $(t, z)$ if and only if $\rho(t', z') \le \rho(t, z)$.

Table 5: Prediction duties under Definition C.1. For a non-row-final cell and for a separator, the predicted token is a cell and its spatial location is expressed through the constant successor offset $\hat{u}_d$. A row-final cell instead predicts a separator whose identity is fixed by position.

| token | predicts | spatial target |
|---|---|---|
| cell $(t, z)$, $z_d < W_d - 1$ | cell $(t, z + \hat{u}_d)$ | $z + \hat{u}_d$ |
| cell $(t, z)$, $z_d = W_d - 1$ | a separator | none |
| separator | first cell of the next row | $\kappa(n) + \hat{u}_d$ |

**Virtual cell index.** We assign a virtual spatial index and time step to every sequence position except the final TIME_SEP. For a cell token $n = \rho(t, z)$, define

$$(\tau(n), \kappa(n)) \coloneqq (t, z).$$

For a separator $n$ whose successor cell at $n + 1$ is $(t', z')$, define

$$(\tau(n), \kappa(n)) \coloneqq (t', z' - \hat{u}_d).$$

Thus, a ROW_SEP inherits its row's time step, while a TIME_SEP assumes the subsequent time step. The final TIME_SEP, at $n = N - 1$, has no successor cell, so $\tau(N - 1)$ and $\kappa(N - 1)$ are left undefined. It poses no query, and as the last position it is visible as a key only to itself. We therefore exclude it from every requirement below which quantifies over virtual addresses, including the gather, and give its column of the positional table arbitrary features; no scored prediction depends on it.

**Lemma C.2** (Constant successor offset)**.** *Under Definition C.1, a cell token $(t, z)$ is followed by a cell token exactly when $z_d < W_d - 1$, in which case that successor is $(t, z + \hat{u}_d)$. Every separator other than the final TIME_SEP is also followed by a cell token. Whenever position $n$ is followed by a cell token, that token, which we call the* target cell *of $n$, is*

$$(\tau(n), \kappa(n) + \hat{u}_d).$$

*Hence every position which predicts a cell does so at the same spatial offset $\hat{u}_d$ from its virtual cell index.*

*Proof.* Within an innermost row, a non-final cell $z$ is followed by $z + \hat{u}_d$, and a row-final cell by its separator. Non-final separators are followed by the first cell of the subsequent row or time step. For a separator with successor $(t', z')$, the virtual index yields $\tau(n) = t'$ and $\kappa(n) = z' - \hat{u}_d$; its target remains $\kappa(n) + \hat{u}_d = z'$. □

*Remark* (Size of the routing set). Because $\mathcal{U} = D_K \cup (D_K + \hat{u}_d)$ is a union of two sets of size $k$, and because the offsets are distinct modulo $(W_1, \ldots, W_d)$ by Definition 1,

$$k \;\le\; |\mathcal{U}| \;\le\; \min(2k, |\mathcal{G}|),$$

which is the bound quoted in §3.2. The lower bound is attained exactly when $D_K + \hat{u}_d = D_K$, in which case $\mathcal{U} = D_K$. This degenerate case is possible: on $\mathbb{Z}_3^2$ with $D_K = \{(0, 0), (0, 1), (0, 2)\}$ and $\hat{u}_d = (0, 1)$, translation by $\hat{u}_d$ permutes $D_K$, so $k = |\mathcal{U}| = 3$. It requires $D_K$ to be a union of cosets of $\langle \hat{u}_d \rangle$, and so to wrap the innermost axis completely; every neighborhood we consider has $|\mathcal{U}| \ge k + 1$.

### C.2 Boundary and separator queries

Serialization introduces two boundary cases: a *row-final* cell ($z_d = W_d - 1$) predicts a separator, and a separator predicts the first cell of the next row. These cases, fixed by emission order, are summarized in Table 5.

**Lemma C.3** (Row-final targets are fixed by position)**.** *Let $n = \rho(t, z)$ hold a cell token with $z_d = W_d - 1$ and $n < N - 1$. Then the token at $n + 1$ is a TIME_SEP if $r(z) = |\mathcal{G}|/W_d - 1$ and a ROW_SEP otherwise. In particular, conditioned on the position $n$, the next token is deterministic and independent of the cellular-automaton kernel $f$.*

*Proof.* The emission order is fixed before any cell state is drawn. Every innermost row is followed by exactly one separator, and the separator is a TIME_SEP exactly for the last row of a time step. Equation (2) therefore determines the next token type from $n$ alone. □

The experimental evaluation (Section 5) scores cell-value targets, excluding separators. Scored predictions include both non-row-final cell queries and separator queries targeting the prediction region. Row-final cell tokens (fraction $1/W_d$) predict separators whose identities are positionally fixed and independent of the rule $f$. However, they remain valid *keys* for Layer 2, as their states are valid observations of $f$ (Remark G.5).

The remaining case is a separator followed by a cell. We chose the virtual index so that this case has the same spatial target convention as an ordinary query.

**Lemma C.4** (Separator queries)**.** *Let $n$ hold a separator with $n < N - 1$, and let $(t', z')$ be the cell token at $n + 1$.*

*(i) If $n$ is a `ROW_SEP` closing row $r$ of time step $t$, then $t' = t$, $r(z') = r + 1$, and $z'_d = 0$. If $n$ is a `TIME_SEP` closing time step $t$, then $(t', z') = (t + 1, 0)$. In either case,*

$$\tau(n) = t', \qquad \kappa(n) = z' - \hat{u}_d, \qquad \kappa(n) + \hat{u}_d = z'.$$

*(ii) If $t' \geq 1$, then every cell in time step $t' - 1 = \tau(n) - 1$ precedes $n$ in the stream. Hence all $k$ parents of the target cell $z'$ are causally visible when the separator predicts it.*

*(iii) If $t' \geq 1$, the parents of $z'$ occur at offsets $D_K$ from $z'$, equivalently at offsets*

$$D_K + \hat{u}_d = D_Q \subseteq \mathcal{U}$$

*from the virtual index $\kappa(n)$. Thus separator queries whose targets are CA-generated cells require no spatial offsets beyond the routing set already defined in Definition 2.*

*Proof.* For part (i), the token following the separator of row $r$ is the first cell of row $r + 1$, while the token following a `TIME_SEP` is the first cell of the next time step. The identities involving $(\tau, \kappa)$ then follow from the definition of the virtual index.

For part (ii), if $n$ is a `ROW_SEP`, then $t' = t$, and all of time step $t - 1$ occurs before every token of time step $t$. If $n$ is a `TIME_SEP`, then $t' = t + 1$, and every cell of time step $t$ appears before its closing separator. The condition $t' \geq 1$ excludes separators inside the initial condition, whose successor cells have no CA parents.

For part (iii), when $t' \geq 1$, the parent set of the target $z' = \kappa(n) + \hat{u}_d$ is $z' + D_K = \kappa(n) + (D_K + \hat{u}_d) = \kappa(n) + D_Q$. □

The two separator types differ only in how their stream position relates to the target time step: a `ROW_SEP` predicts a cell in its own time step (reading parents from $t - 1$), whereas a `TIME_SEP` predicts the first cell of the next step (reading parents from $t$). Indexing by $\tau$ unifies these cases. Initial-condition separators predicting initial states fall outside the transition-prediction task. The transformer constructions (Appendix G) use separators as queries when predicting CA cells, gating them out as keys via token content; this requires no extra spatial offsets or positional coordinates.

### C.3 Limits of displacement-only positional biases

In this section, we bound what a positional bias depending only on sequence displacement can do. No fixed scalar bias scores every desired routing cell strictly above all other causally visible cells at every query position; periodic wraparound prevents this one-layer isolation (Theorem C.6).

The absolute positional table (Section 3.3) lacks an explicit coordinate system, requiring spatial structure to be learned from data. A simpler alternative is a scalar additive bias depending solely on sequence displacement (Raffel et al., 2020; Press et al., 2021).

On an open grid, a fixed spatial offset yields a constant sequence displacement between any two valid cells, provided no boundary truncation or padding changes the row length. On a torus, wraparound disrupts this consistency. Theorem C.6 identifies two queries where conflicting displacements prevent any fixed displacement-only bias from isolating the desired cells.

For the formulas below, write

$$R \coloneqq W_d + 1, \qquad S \coloneqq (|\mathcal{G}|/W_d)R, \qquad r(z) \coloneqq \sum_{j<d} z_j \prod_{j<l<d} W_l.$$

Here $R$ is the row stride, $S$ is the time-step stride, and $r(z)$ agrees with Definition C.1. Equation (2) becomes

$$\rho(t, z) = tS + r(z)R + z_d. \tag{3}$$

**Definition C.5** (Sequence-relative bias)**.** A *sequence-relative bias* is a function

$$b : \{0, \dots, N-1\} \to \mathbb{R}$$

whose contribution to the attention score between a query at position $n$ and a key at position $n' \le n$ depends only on the displacement $n - n'$. Separator keys may additionally be gated out by token content.

For the isolation property below, ordinary cell keys are ranked solely by the positional score $b(n-n')$. For a non-row-final cell query $(t, z)$ with $1 \le t < T$ and $z_d < W_d - 1$, such a bias *isolates* an offset set $H \subseteq \mathcal{G}$ if every desired cell $(t-1, z+\delta)$ with $\delta \in H$ receives a strictly larger positional score than every causally visible cell outside this desired set.

Allowing separator keys to be gated only strengthens the class in Definition C.5. The witnesses below compare cell tokens with cell tokens and never use a separator as a competing key.

*Example* (The elementary one-dimensional contradiction)*.* For the elementary neighborhood, $D_K = \{-1, 0, 1\}$ and $\mathcal{U} = \{-1, 0, 1, 2\}$. Let $W \ge 4$. At the query $(2, 0)$, the desired cell at offset $-1$ wraps to the right edge of the previous time step and lies at sequence displacement 2. The cell at displacement $2W - 1$ is visible but lies two time steps back, so correct isolation requires

$$b(2) > b(2W - 1).$$

At the query $(t, W-2)$, the desired cell at offset 2 wraps to the left edge of the previous time step and lies at displacement $2W - 1$, while displacement 2 points to a visible cell in the current time step. Correct isolation now requires

$$b(2W - 1) > b(2).$$

The two inequalities are incompatible. Theorem C.6 generalizes this argument.

*Remark* (What the theorem does not cover)*.* A bias which may also read the query position modulo the row stride is already grid-indexed: $n \bmod R = z_d$ on cell tokens, which exposes exactly the boundary information used by the contradiction. Standard one-dimensional rotary embeddings (Su et al., 2024) enter the logits through a position-dependent bilinear form, so they too fall outside Definition C.5 and the theorem. We do not analyze either class here.

**Theorem C.6** (Sequence-relative obstruction)**.** *Fix the serialization of Definition C.1 with $T \ge 3$.*

*(i) Suppose $d = 1$, $W \ge 4$, and $\mathcal{U}$ contains some negative offset $-m$ with $1 \le m \le W - 3$ and some positive offset $p$ with $2 \le p \le W - 2$. Then no single sequence-relative bias isolates $\mathcal{U}$ for every query cell $z < W - 1$.*

*(ii) Suppose $d \ge 2$, $W_d \ge 2$, and some $\delta \in D_K$ has a nonzero component modulo the corresponding side length on an outer axis; that is, $\delta_j \not\equiv 0 \pmod{W_j}$ for some $j < d$. Then no partition*

$$\mathcal{U} = H_1 \sqcup \cdots \sqcup H_h$$

*together with sequence-relative biases $b_1, \dots, b_h$ can make each $b_a$ isolate $H_a$ at every query. In particular, no single sequence-relative bias isolates $\mathcal{U}$.*

Part (i) includes the elementary neighborhood by taking $m = 1$ and $p = 2$. Part (ii) applies whenever the neighborhood is not confined to one innermost row, including the usual von Neumann and Moore neighborhoods in dimension at least two. The result concerns one-layer isolation in the sense of Definition C.5; it does not rule out a model which assembles the routing set through several layers or which is supplied with additional spatial information.

*Proof. Part (i).* Write $W = W_1$. Then $R = S = W + 1$ and $\rho(t, z) = (W+1)t + z$. For a query $(t, z)$, the displacement to the desired cell $(t-1, z+\delta)$ is

$$\rho(t, z) - \rho(t-1, z+\delta) = \begin{cases} (W+1) - \delta, & 0 \le z + \delta < W, \\ (2W+1) - \delta, & z + \delta \ge W, \\ 1 - \delta, & z + \delta < 0, \end{cases}$$

where the last two cases use toroidal wraparound.

Take first the query $(2, 0)$, which $T \geq 3$ and $W \geq 4$ make a non-row-final cell query in the sense of Definition C.5. The desired cell at offset $-m$ wraps below and therefore lies at displacement $m + 1$. The position at displacement $2W + 1 - p$ is

$$\rho(2, 0) - (2W + 1 - p) = p + 1,$$

which is the visible cell $(0, p + 1)$ because $p \leq W - 2$. It lies in time step 0, not in the desired parent step 1, so it is not one of the routed cells. Isolation therefore requires

$$b(m + 1) > b(2W + 1 - p). \tag{4}$$

Now take the query $(t, W - 2)$ for any $1 \leq t < T$, which is a non-row-final cell query because $W \geq 4$. The desired cell at offset $p$ wraps above and lies at displacement $2W + 1 - p$. The position at displacement $m + 1$ is

$$\rho(t, W - 2) - (m + 1) = \rho(t, W - 3 - m),$$

which is a cell in the current time step because $m \leq W - 3$. It precedes the query and is therefore visible, but it does not belong to the previous time step. Hence isolation requires

$$b(2W + 1 - p) > b(m + 1), \tag{5}$$

contradicting (4). This proves part (i).

*Part (ii).* Let $M_{\mathrm{row}} \coloneqq |\mathcal{G}|/W_d$ be the number of rows in one time step, so $S = M_{\mathrm{row}}R$, and write

$$P_j \coloneqq \prod_{j<l<d} W_l, \qquad r(z) = \sum_{j<d} z_j P_j.$$

Choose $\delta \in D_K$ such that $\delta_j \not\equiv 0 \pmod{W_j}$ for some $j < d$, and represent each coordinate by its residue $\delta_j \in \{0, \ldots, W_j - 1\}$. Let $i < d$ be the smallest index with $\delta_i \neq 0$, and define

$$C \coloneqq \sum_{j<d} \delta_j P_j, \qquad C' \coloneqq \sum_{i<j<d} \delta_j P_j, \qquad G \coloneqq W_i P_i R, \qquad A \coloneqq S - CR - \delta_d.$$

Because $i$ is the first nonzero outer coordinate, mixed-radix bounds give $C' < P_i$ and hence

$$P_i \leq C < W_i P_i \leq M_{\mathrm{row}}.$$

Consider first the query $(2, 0)$, which $T \geq 3$ and $W_d \geq 2$ make a non-row-final cell query. The desired routed cell at offset $\delta$ is $(1, \delta)$, whose sequence displacement is

$$\rho(2, 0) - \rho(1, \delta) = 2S - (S + CR + \delta_d) = A.$$

The position at displacement $A + G$ is

$$2S - A - G = S + (C - W_i P_i)R + \delta_d = (M_{\mathrm{row}} + C - W_i P_i)R + \delta_d.$$

Since $0 \leq M_{\mathrm{row}} + C - W_i P_i < M_{\mathrm{row}}$, this is a cell in time step 0. It is visible, but all desired routed cells for the query lie in time step 1, so it is not a desired key. Therefore the bias assigned to any part containing $\delta$ must satisfy

$$b(A) > b(A + G). \tag{6}$$

For the reverse inequality, take

$$z^* \coloneqq (W_1 - 1, \ldots, W_{i-1} - 1, W_i - \delta_i, 0, \ldots, 0)$$

at time $t = 1$. Since $z^*_d = 0$ and $W_d \geq 2$, this is a non-row-final cell query. Since

$$\sum_{j<i} (W_j - 1) P_j = M_{\mathrm{row}} - W_i P_i,$$

we have $r(z^*) = M_{\mathrm{row}} - \delta_i P_i$. Adding $\delta$ wraps coordinate $i$ to zero and gives

$$r(z^* + \delta) = M_{\mathrm{row}} - W_i P_i + C'.$$

Thus the desired cell $(0, z^* + \delta)$ lies at displacement

$$\begin{aligned}\rho(1, z^*) - \rho(0, z^* + \delta) &= S + \big(W_i P_i - \delta_i P_i - C'\big)R - \delta_d \\ &= A + G.\end{aligned}$$

By contrast, the position at displacement $A$ is

$$\rho(1, z^*) - A = S + C'R + \delta_d,$$

a cell in time step 1, in row $C'$ and column $\delta_d$. It precedes the query because

$$C'R + \delta_d < (C' + 1)R \le P_i R \le (M_{\text{row}} - \delta_i P_i)R,$$

and it is not a desired routed cell because the desired cells lie in time step 0. Hence the same bias must satisfy

$$b(A + G) > b(A), \tag{7}$$

contradicting (6).

This contradiction applies to whichever partition part contains $\delta$. Therefore no family of sequence-relative biases, one per part of a partition of $\mathcal{U}$, can isolate all parts. This proves part (ii). □

**Spatial relative positions and learned routing.** A spatial relative-position bias in $d$ dimensions, augmented with temporal displacement, can avoid this obstruction. For ordinary cell tokens, such a bias can take the form

$$b_{\text{sp}}(t - t',\, z' - z),$$

where $z' - z \in \mathcal{G}$ is computed coordinatewise modulo $(W_1, \ldots, W_d)$. Assigning a higher score when $t - t' = 1$ and $z' - z \in \mathcal{U}$ than otherwise isolates the desired routing cells. The same spatial offset therefore remains identifiable across periodic boundaries.

This parameterization explicitly supplies the mapping from sequence positions to spatial and temporal coordinates, together with the periodic boundary structure, even when the bias values themselves are learned. Our experimental models instead use freely learned absolute position embeddings indexed only by sequence position, without an explicit coordinate-based parameterization. The serialization and separators supply structural cues, but the model must learn how to use them for spatial routing. Our constructions establish that such routing can be represented, and our experiments provide evidence that gradient-based training learns routing behavior consistent with these constructions.

# D Coverage and Bayes-optimal prediction

All distributional guarantees apply to *resolvable* queries and are measured against the *Bayes-optimal predictor*. Section D.1 characterizes coverage and specifies the queries excluded by resolvability, where optimal prediction relies solely on position and the prior. Section D.2 derives the Bayes-optimal predictor (Theorem 3). These results rely exclusively on Definition 1 and the serialization from Appendix C, without reference to transformer architectures.

## D.1 Coverage and resolvable queries

We adapt Definition 4 to the serialized stream to address two issues: the parent configuration $P_{t,z}$ (Definition 1) is defined only for $t \ge 1$ (as the initial step is drawn independently), and the first cell of every row is predicted at a separator token (Lemma C.4), which is scored alongside interior cells (Section 5).

**Querying positions.** A position $n$ *poses a query* when it has a target cell in the sense of Lemma C.2 and that target cell lies in a time step $\tau(n) \ge 1$. Two families qualify, and they are the first and third rows of Table 5:

(a) a cell token $(t, z)$ with $t \ge 1$ and $z_d < W_d - 1$, whose target cell is $z + \hat{u}_d$ in its own time step;

(b) a separator at $n < N - 1$ with $\tau(n) \ge 1$, whose target cell is $\kappa(n) + \hat{u}_d$, the first cell of the next row, again in step $\tau(n)$.

In both cases the query configuration is

$$Q_n \;\coloneqq\; P_{\tau(n),\,\kappa(n)+\hat{u}_d},$$

which is defined because $\tau(n) \geq 1$, and the target is a draw from $f(\cdot \mid Q_n)$. For a cell token this is the $Q_{t,z}$ of Definition 4. Theorem H.9 gives the separator case of the guarantee, under the same definition.

**Counting positions.** The counts range over the cells generated by $f$:

$$n_c \;\coloneqq\; \#\big\{(t', z') \;:\; t' \geq 1,\; \rho(t', z') \leq n,\; P_{t',z'} = c\big\}, \qquad n_{c,y} \;\coloneqq\; \#\{\text{those with } x_{t',z'} = y\}. \tag{8}$$

Above, $c \in \mathcal{S}^k$ is a configuration and $n$ is the sequence position posing the query. We need the restriction $t' \geq 1$ for two reasons: $P_{0,z'}$ is undefined, and $x_{0,z'}$ is an initial state, so it says nothing about the kernel. A configuration $c$ is *covered* at $n$ when $n_c \geq 1$, and the query at $n$ is *resolvable* when its configuration $Q_n$ is covered.

Coverage is a property of the stream alone, and it is monotone in position.

**Lemma D.1** (Coverage is monotone)**.** *Fix a trajectory and a configuration $c \in \mathcal{S}^k$.*

*(i) $n_c$, evaluated at $(t, z)$, is non-decreasing in $\rho(t, z)$.*

*(ii) If a query posing $c$ is resolvable at $(t, z)$, then every query posing $c$ at a position $(t'', z'')$ with $\rho(t'', z'') \geq \rho(t, z)$ is resolvable.*

*(iii) A query is resolvable if and only if $n_c \geq 1$ for its own configuration $c = Q_{t,z}$.*

*Proof.* Since $\rho$ is an order-preserving injection (Definition C.1), the sets $A(n) \coloneqq \{(t', z') : t' \geq 1,\; \rho(t', z') \leq n\}$ form a chain under inclusion. The condition $t' \geq 1$ excludes the initial time step uniformly, preserving the chain structure. Because $n_c$ at $(t, z)$ counts elements of $A(\rho(t, z))$ with parent configuration $c$, it is non-decreasing along this chain, proving (i). Statement (ii) follows from (i) and (iii). For (iii), $c$ is covered at $(t, z)$ if and only if $A(\rho(t, z))$ contains an element with parent configuration $c$, meaning $n_c \geq 1$. □

Part (iii) gives a nonempty matched set. Both the deterministic and the stochastic analysis divide by $n_c$, so both need resolvability.

Evaluating only on resolvable queries excludes three classes of cell token.

**Lemma D.2** (What resolvability excludes)**.** *A cell token either poses a resolvable query or falls into exactly one of the following three classes, tested in this order.*

*(i)* Initial-step cells*, those with $t = 0$. Their target is a separator when $z_d = W_d - 1$ and an initial state otherwise. The separator is determined by the position $n$, and the initial state is drawn from the initial-state distribution of Definition 1, so neither depends on $f$.*

*(ii)* Row-final cells*, those with $t \geq 1$ and $z_d = W_d - 1$. Their target is a separator, determined by the position $n$ alone, so its conditional law given everything observed is a point mass and carries zero mutual information with $f$.*

*(iii)* Novel configurations*, those with $t \geq 1$ and $z_d < W_d - 1$ whose query configuration has $n_c = 0$. There the Bayes posterior predictive of Theorem D.3 is the uniform distribution $\widehat{P}_\alpha(y \mid c) = 1/V$ for every $y$, whatever $f$ is.*

*In all three classes the optimal prediction depends on the position and the prior alone, and the trajectory adds nothing.*

*Proof.* Tokens with $t = 0$ fall into class (i); those with $t \geq 1$ and $z_d = W_d - 1$ into class (ii); and those with $t \geq 1$, $z_d < W_d - 1$, and an uncovered configuration into class (iii) (Lemma D.1(iii)). The three tests are applied in that order, so the classes are disjoint, and all other cell tokens pose resolvable queries. For (i), an initial-step row-final cell predicts a positionally fixed separator (Lemma C.3) and every other initial-step cell predicts an initial state, drawn independently of $f$. For (ii), a row-final cell predicts a deterministically positioned separator (Lemma C.3), conveying zero mutual information

about $f$. For (iii), substituting $n_c = n_{c,y} = 0$ into Theorem D.3 yields $\widehat{P}_\alpha(y \mid c) = 1/V$, the mean of the $\mathrm{Dir}(\alpha \mathbf{1}_V)$ prior, independent of the trajectory. □

*Remark.* The excluded classes differ in frequency and rationale. Initial-step cells (fraction $1/T$) and novel configurations are excluded because their Bayes-optimal prediction is uninformative about the inferred rule, scoring at $1/V$. Row-final cells (fraction $1/W_d$) are universally excluded because they deterministically predict a separator; scoring them would artificially inflate accuracy. The novel configurations become rare as $T$ grows. Once an uncovered configuration is queried, its target is observed and becomes a precedent, and by Lemma D.1(ii) it stays covered. Each of the $V^k$ configurations therefore yields at most one uncovered query in the whole stream, so for fixed $V$, $k$ and $\mathcal{G}$ their number is bounded independently of $T$ and their fraction vanishes. Consequently, we rejection-sample evaluation trajectories to ensure every scored query is resolvable (Appendix I.3).

### D.2 Bayes-optimal posterior prediction

Under the prior of Definition 1, no predictor attains lower expected log loss than the posterior predictive, which relies on counts (Theorem D.3). We measure divergences against this Bayes-optimal predictor, and Theorem 9 bounds the circuit's divergence from it. For next-cell accuracy the Bayes act is the $\arg\max$ of the posterior predictive. Sampling from it is a different predictor: at predictive distribution $(0.9, 0.1)$, a sample matches the target with probability $0.82$, while the mode matches with probability $0.9$.

**Theorem D.3** (Bayes posterior predictive; Theorem 3 of the body)**.** *Let $f$ be drawn as in Definition 1 and let a trajectory be serialized as in Appendix C. For every token with resolvable query $c$ and every $y \in \mathcal{S}$, the posterior predictive distribution of the target given everything observed up to that token is exactly*

$$\widehat{P}_\alpha(y \mid c) \;=\; \frac{n_{c,y} + \alpha}{n_c + V\alpha},$$

*and no predictor attains smaller expected log-loss.*

*Proof.* By Definition 1, cells at $t \geq 1$ are conditionally independent across $z$ given the previous step, and the initial time step is drawn independently of $f$. The trajectory likelihood factorizes into an initial-step term and one multinomial factor per configuration $c$, with counts $n_{c,y}$. The initial-step factor $p(x_{0,\cdot})$ is independent of $f$ and cancels from the posterior. The independent Dirichlet prior is conjugate, yielding the posterior $\bigotimes_c \mathrm{Dir}(\alpha \mathbf{1} + n_{c,\cdot})$. The mean of a $\mathrm{Dir}(\beta)$ distribution is $\beta_y / \sum_v \beta_v$, producing the stated formula. Under log-loss, the Bayes act is exactly the posterior predictive. □

*Remark.* The proof does not use resolvability, and the formula remains valid at $n_c = 0$, where Lemma D.2(iii) shows it returns the prior mean. We state resolvability in the theorem because it is the regime where the formula is informative, and because every later result which uses this predictor divides by $n_c$.

**Corollary D.4** (Empirical-frequency limit and deterministic lookup)**.** *As $\alpha \to 0$ the predictor of Theorem D.3 tends to the empirical frequency $n_{c,y}/n_c$. If in addition every $f(\cdot \mid c)$ is a point mass, meaning the deterministic corner of Definition 1, then a single observation of $c$ determines its image, so the predictor is the lookup table and is* exact *on every resolvable query.*

*Proof.* The limit is immediate from the displayed formula. Under a point mass $f(\cdot \mid c) = \mathbf{1}_{y^\star}$ every observation of $c$ has target $y^\star$, so $n_{c,y}/n_c = \mathbf{1}[y = y^\star]$ as soon as $n_c \geq 1$, which Lemma D.1(iii) supplies at a resolvable query. □

*Remark.* Corollary D.4 characterizes the *predictor*. In the deterministic setting the Bayes-optimal predictor is exact, so exact-match accuracy is a meaningful metric and $100\%$ is attainable. Theorem H.9*(ii)* characterizes the *circuit*, and it shows that the two-layer construction attains that floor once the score gap is large enough.

Outside the deterministic regime the Bayes-optimal predictor sits strictly below $100\%$ accuracy, because the query configuration determines only a distribution over the target. Table 1 therefore reports the Kullback–Leibler divergence from the $k$-gram estimate of $\widehat{P}_\alpha$, and Theorem 9 bounds the divergence from $\widehat{P}_\alpha$ itself. A predictor which matches the Bayes-optimal predictor still errs on a strictly positive fraction of stochastic queries.

# PART II — SPATIAL ROUTING PRIMITIVES

## E SPATIAL ADDRESSING WITH GRID-INDEPENDENT WIDTH

Appendix C demonstrates that a sequence-relative additive bias (Definition C.5) cannot route a toroidal neighborhood. Consequently, the Layer-1 positional table is indexed spatially. This section derives the required table width. Lemma 6 establishes that a positional bias can isolate any target offset set using a grid-independent number of frequencies; the table width is this frequency count plus two.

The section is structured as follows. Section E.1 shows the coordinate dimension of a translation-invariant bias equals its Fourier spectral support size. Section E.2 bounds this support for isolating kernels, which strictly order a target set above its complement. Section E.3 translates this ordering into an attention mass bound. Finally, Section E.4 constructs the positional table and its projections, establishing the interface for Appendix G, while Section E.5 provides a numerical example.

### E.1 FOURIER REPRESENTATION OF SPATIAL ATTENTION SCORES

We apply harmonic analysis on finite abelian groups (see Terras, 1999, §1.2 and §1.10) without proof.

The grid $\mathcal{G} = [W_1] \times \cdots \times [W_d]$ is the product of the cyclic groups $\mathbb{Z}_{W_1}, \ldots, \mathbb{Z}_{W_d}$ under coordinate-wise addition, and has order $|\mathcal{G}| = \prod_j W_j$. Its dual $\widehat{\mathcal{G}}$ is the group of characters, indexed by the same set: for $\omega \in \widehat{\mathcal{G}}$ and $\delta \in \mathcal{G}$ put

$$\varphi_\omega(\delta) \;\coloneqq\; 2\pi \sum_{j=1}^{d} \frac{\omega_j \delta_j}{W_j}, \qquad \chi_\omega(\delta) \;\coloneqq\; e^{i\varphi_\omega(\delta)}. \tag{9}$$

Two consequences of (9) are frequently used. First, the phase is additive,

$$\varphi_\omega(\delta + \delta') \;\equiv\; \varphi_\omega(\delta) + \varphi_\omega(\delta') \pmod{2\pi}, \qquad \text{equivalently} \qquad \chi_\omega(\delta + \delta') \;=\; \chi_\omega(\delta)\, \chi_\omega(\delta'), \tag{10}$$

meaning a displacement in the argument rotates the value. The congruence is needed because $\delta + \delta'$ is reduced modulo the side lengths; the character identity, and every sine and cosine computed from it, hold exactly. Second, characters are orthonormal:

$$\mathop{\mathbb{E}}_{\omega \sim \mathrm{Unif}(\widehat{\mathcal{G}})} \chi_\omega(y) \;=\; \frac{1}{|\mathcal{G}|} \sum_{\omega \in \widehat{\mathcal{G}}} \chi_\omega(y) \;=\; \begin{cases} 1 & y = 0, \\ 0 & y \neq 0, \end{cases} \qquad y \in \mathcal{G}. \tag{11}$$

Since the characters form an orthonormal basis of $\mathbb{C}^{\mathcal{G}}$, every $g : \mathcal{G} \to \mathbb{C}$ has a unique expansion $g = \sum_\omega \hat{g}(\omega)\chi_\omega$. The *spectrum* of $g$ is $\operatorname{supp} \hat{g} \coloneqq \{\omega : \hat{g}(\omega) \neq 0\}$, and it is conjugate-symmetric, $\hat{g}(-\omega) = \overline{\hat{g}(\omega)}$, when $g$ is real. We write $e_1, \ldots, e_d$ for the standard generators of $\widehat{\mathcal{G}}$, and

$$d_S \;\coloneqq\; \big|\operatorname{supp} \hat{g} \setminus \{0\}\big|$$

for the number of nontrivial characters in the spectrum, the *addressing dimension* of $g$.

A displacement-dependent positional bias $z - z'$ induces a $\mathcal{G}$-circulant score matrix. Lemma E.1 establishes its minimal rank factorization, with transformer implications discussed in §E.4.

**Lemma E.1** (Realization of translation-invariant spatial biases)**.** *Let $g : \mathcal{G} \to \mathbb{R}$ and $\Omega \coloneqq \operatorname{supp} \hat{g}$. The least $n$ for which there exist $p : \mathcal{G} \to \mathbb{R}^n$ and $Q \in \mathbb{R}^{n \times n}$ with*

$$p_z^\top Q\, p_{z'} \;=\; g(z - z') \qquad \textit{for all } z, z' \in \mathcal{G}$$

*is exactly $|\Omega|$. A $Q$ attaining it may be taken block diagonal with blocks of size $1$ and $2$, the $2 \times 2$ blocks being scaled rotations.*

*Proof. The score matrix and its eigenvalues.* Let $M$ be the matrix indexed by $\mathcal{G} \times \mathcal{G}$ with entries $M_{z,z'} \coloneqq g(z - z')$. It is constant along the orbits of the diagonal action of $\mathcal{G}$ on itself, making it a $\mathcal{G}$-circulant matrix, the $d$-dimensional analogue of a circulant. Every character is an eigenvector of $M$: substituting $y \coloneqq z - z'$ and using (10),

$$\begin{aligned}(M\chi_\omega)(z) \;&=\; \sum_{z' \in \mathcal{G}} g(z - z')\, \chi_\omega(z') \;=\; \sum_{y \in \mathcal{G}} g(y)\, \chi_\omega(z - y) \\ &=\; \chi_\omega(z) \sum_{y \in \mathcal{G}} g(y)\, \overline{\chi_\omega(y)} \;=\; |\mathcal{G}|\, \hat{g}(\omega)\, \chi_\omega(z).\end{aligned} \tag{12}$$

The $|\mathcal{G}|$ characters are a basis, so (12) diagonalizes $M$ and its eigenvalues are $\{|\mathcal{G}|\hat{g}(\omega)\}_{\omega \in \widehat{\mathcal{G}}}$. Exactly $|\Omega|$ of them are nonzero, so $\operatorname{rank} M = |\Omega|$.

*Lower bound.* Suppose $p$ and $Q$ are as in the statement and let $P$ be the $n \times |\mathcal{G}|$ matrix whose columns are the vectors $p_z$. Then $M = P^\top Q P$, so

$$|\Omega| \;=\; \operatorname{rank} M \;\le\; \operatorname{rank} Q \;\le\; n.$$

*Construction.* For the converse, partition $\Omega$ into orbits of the involution $\omega \mapsto -\omega$. Because $g$ is real its spectrum is conjugate-symmetric, so collecting each orbit into a single real term gives

$$g(\delta) \;=\; \sum_{\text{orbits } \{\omega, -\omega\}} \big[\alpha_\omega \cos \varphi_\omega(\delta) + \beta_\omega \sin \varphi_\omega(\delta)\big], \tag{13}$$

with $\alpha_\omega, \beta_\omega \in \mathbb{R}$. An orbit is *self-paired* when $\omega = -\omega$, that is $2\omega = 0$; there $\varphi_\omega$ takes values in $\{0, \pi\}$, so $\sin \varphi_\omega \equiv 0$ and $\beta_\omega = 0$.

Assign coordinates as follows: a self-paired orbit, which contributes one element to $\Omega$, receives the single coordinate $\cos \varphi_\omega(z)$; a conjugate pair, which contributes two, receives the two coordinates

$$e_\omega(z) \;\coloneqq\; \begin{bmatrix}\cos \varphi_\omega(z) \\ \sin \varphi_\omega(z)\end{bmatrix} \;\in\; \mathbb{R}^2. \tag{14}$$

The total is $|\Omega|$ coordinates. Take $Q$ block diagonal with the blocks

$$\big[\,\alpha_\omega\,\big] \quad \text{on a self-paired orbit,} \qquad M_\omega \;\coloneqq\; \begin{bmatrix}\alpha_\omega & -\beta_\omega \\ \beta_\omega & \alpha_\omega\end{bmatrix} \quad \text{on a conjugate pair.} \tag{15}$$

It remains to check that each block contributes its term of (13). Write $A \coloneqq \varphi_\omega(z)$ and $B \coloneqq \varphi_\omega(z')$. Expanding the quadratic form and applying the angle-difference identities $\cos(A - B) = \cos A \cos B + \sin A \sin B$ and $\sin(A - B) = \sin A \cos B - \cos A \sin B$,

$$\begin{aligned}e_\omega(z)^\top M_\omega\, e_\omega(z') &= [\cos A \quad \sin A] \begin{bmatrix}\alpha_\omega & -\beta_\omega \\ \beta_\omega & \alpha_\omega\end{bmatrix} \begin{bmatrix}\cos B \\ \sin B\end{bmatrix} \\ &= \alpha_\omega\big(\cos A \cos B + \sin A \sin B\big) + \beta_\omega\big(\sin A \cos B - \cos A \sin B\big) \\ &= \alpha_\omega \cos(A - B) + \beta_\omega \sin(A - B).\end{aligned} \tag{16}$$

By (10), $A - B = \varphi_\omega(z) - \varphi_\omega(z') = \varphi_\omega(z - z')$, so (16) is exactly the $\omega$-term of (13) evaluated at $z - z'$. The self-paired blocks reduce to the same computation with the sine coordinate absent. Summing over the orbits gives $p_z^\top Q\, p_{z'} = g(z - z')$.

Finally, $M_\omega = r_\omega R(\theta_\omega)$ with $r_\omega = \sqrt{\alpha_\omega^2 + \beta_\omega^2}$, $\theta_\omega = \operatorname{atan2}(\beta_\omega, \alpha_\omega)$ and $R(\theta)$ the rotation by $\theta$, so the $2 \times 2$ blocks are scaled rotations, as claimed. □

*Remark.* The lower bound establishes the addressing dimension $d_S$ (§E.1) as an intrinsic property of the positional bias. In §E.4, the scaled-rotation construction (15) allows an attention head to select its routing offset via a linear map applied to a shared positional table, minimizing parameter count.

### E.2 ISOLATING A SET OF OFFSETS

**Definition E.2** (Isolation)**.** Let $\emptyset \neq \mathcal{N} \subsetneq \mathcal{G}$ and $\gamma > 0$. A function $g : \mathcal{G} \to \mathbb{R}$ *isolates* $\mathcal{N}$ *with margin* $\gamma$ if

$$\min_{\delta \in \mathcal{N}} g(\delta) \;-\; \max_{\delta \notin \mathcal{N}} g(\delta) \;\geq\; \gamma.$$

Isolation is an ordering condition. Since additive constants leave the softmax invariant, $g$ isolates $\mathcal{N}$ if and only if $g + \mathrm{const}$ does. We therefore discard the trivial character $\omega = 0$ and report counts for $\operatorname{supp} \hat{g} \setminus \{0\}$.

The elementary construction multiplies one factor per value the target set assumes on each axis. It vanishes exactly on the bounding box of $\mathcal{N}$ and is strictly positive elsewhere, reducing the problem to the box. When the box coincides with the target set, it yields an exact solution (e.g., Construction I in Appendix G).

**Lemma E.3** (Box kernels)**.** *Let $\emptyset \neq \mathcal{N} \subseteq \mathcal{G}$, let $V_j \coloneqq \{u_j : u \in \mathcal{N}\}$ be the projection of $\mathcal{N}$ to axis $j$, and put*

$$P_j(x) \;\coloneqq\; \prod_{v \in V_j} \Big(1 - \cos\big(2\pi(x_j - v)/W_j\big)\Big), \qquad Q \;\coloneqq\; \sum_{j=1}^{d} P_j, \qquad B \;\coloneqq\; V_1 \times \cdots \times V_d.$$

*Then $Q \geq 0$ vanishes exactly on $B$, which contains $\mathcal{N}$ and satisfies $|B| \leq |\mathcal{N}|^d$, and*

$$\operatorname{supp} \hat{Q} \setminus \{0\} \;\subseteq\; \{c\, e_j : j \leq d,\; 1 \leq |c| \leq |\mathcal{N}|\},$$

*at most $2d|\mathcal{N}|$ characters. Consequently:*

- *(i) if $B = \mathcal{N} \neq \mathcal{G}$, then $-Q$ isolates $\mathcal{N}$ with margin $\varepsilon \coloneqq \min_{x \notin B} Q(x) > 0$ at no more than $2d|\mathcal{N}|$ nontrivial frequencies;*
- *(ii) a single offset, $\mathcal{N} = \{\delta\}$, is that case: $B = \{\delta\}$, $Q(x) = \sum_j \big(1 - \cos(2\pi(x_j - \delta_j)/W_j)\big)$ is of degree one on each axis, and $\{\delta\}$ is isolated at* at most $2d$ *frequencies, the pair* $(\cos, \sin)$ *on each axis, with margin $1 - \cos(2\pi/W_{\max})$, where $W_{\max} \coloneqq \max_j W_j$ is the largest side length. The count is exactly $2d$ when every $W_j \geq 3$; an axis with $W_j = 2$ contributes one character instead of two, since there $e_j = -e_j$ and the orbit is self-paired.*

*Proof. Non-negativity and the zero set.* Each factor of $P_j$ has the form $1 - \cos\theta \in [0, 2]$ and vanishes exactly when $\theta \in 2\pi\mathbb{Z}$, that is when $x_j = v$ in $\mathbb{Z}_{W_j}$. So $P_j \geq 0$, and $P_j(x) = 0$ if and only if $x_j \in V_j$. A sum of non-negative terms vanishes only where every term does, so $Q \geq 0$ and

$$Q(x) = 0 \quad\Longleftrightarrow\quad x_j \in V_j \text{ for every } j \quad\Longleftrightarrow\quad x \in B.$$

*The box contains the target set, and is not much larger.* If $u \in \mathcal{N}$ then $u_j \in V_j$ for every $j$ by the definition of $V_j$, so $\mathcal{N} \subseteq B$. Each $V_j$ is the image of $\mathcal{N}$ under a coordinate projection, so $|V_j| \leq |\mathcal{N}|$ and $|B| = \prod_j |V_j| \leq |\mathcal{N}|^d$.

*The spectrum.* Fix $j$. Each factor of $P_j$ is a function of $x_j$ alone whose spectrum lies in $\{0, \pm e_j\}$, a trigonometric polynomial of degree 1 in $x_j$. The product-to-sum identity

$$\cos(a)\cos(b) \;=\; \tfrac{1}{2}\big[\cos(a - b) + \cos(a + b)\big]$$

and its two companions show that multiplying a trigonometric polynomial of degree $m$ by one of degree 1 yields one of degree at most $m + 1$. By induction on the $|V_j|$ factors, $P_j$ is a trigonometric polynomial in $x_j$ of degree at most $|V_j| \leq |\mathcal{N}|$, so

$$\operatorname{supp} \hat{P}_j \;\subseteq\; \{c\, e_j : |c| \leq |\mathcal{N}|\}.$$

Summing over $j$ gives the stated inclusion for $Q$. Discarding $\omega = 0$, each axis contributes at most the $2|\mathcal{N}|$ characters $\pm e_j, \pm 2e_j, \ldots, \pm|\mathcal{N}|e_j$, for at most $2d|\mathcal{N}|$ in all.

*Part (i).* Suppose $B = \mathcal{N}$. Then $Q$ vanishes on $\mathcal{N}$, and off $\mathcal{N}$ it is positive; since $\mathcal{G}$ is finite, $\varepsilon = \min_{x \notin B} Q(x)$ is a minimum over a finite set on which $Q$ does not vanish, hence $\varepsilon > 0$. Therefore

$$-Q(x) = 0 \quad (x \in \mathcal{N}), \qquad -Q(x) \leq -\varepsilon \quad (x \notin \mathcal{N}),$$

which is isolation with margin $\varepsilon$. The frequency count is the one just established, and negation does not change a spectrum.

*Part (ii).* Let $\mathcal{N} = \{\delta\}$. Then $V_j = \{\delta_j\}$ is a single value for each $j$, so each $P_j$ consists of one factor, $B = V_1 \times \cdots \times V_d = \{\delta\}$, and

$$Q(x) \;=\; \sum_{j=1}^{d} \Big(1 - \cos\big(2\pi(x_j - \delta_j)/W_j\big)\Big).$$

Each summand has spectrum $\{0, \pm e_j\}$, so $\operatorname{supp}\hat{Q} \setminus \{0\}$ is contained in $\{\pm e_j\}_{j \le d}$, one $(\cos, \sin)$ pair per axis and hence at most $2d$ elements. The containment is an equality when every $W_j \ge 3$. If $W_j = 2$ then $2e_j = 0$ in $\widehat{\mathcal{G}}$, so $e_j$ and $-e_j$ are the same character, $\varphi_{e_j}$ takes values in $\{0, \pi\}$, and that axis contributes a single self-paired character, so the count drops by one. Lemma E.1 then assigns that axis one coordinate, so the realized width matches the reduced count. Padding to $2d$ recovers a single uniform formula where one is preferred. Since $B = \mathcal{N}$, part (i) applies, and it remains to bound $\varepsilon$ from below. Let $x \ne \delta$. Then $x_j \ne \delta_j$ for some axis $j$, so $x_j - \delta_j$ is a nonzero residue modulo $W_j$ and its phase $2\pi(x_j - \delta_j)/W_j$ is at angular distance at least $2\pi/W_j$ from 0. Cosine is decreasing in the distance to 0 on $[0, \pi]$, so that summand is at least $1 - \cos(2\pi/W_j)$, and the remaining summands are non-negative:

$$Q(x) \;\ge\; 1 - \cos\big(2\pi/W_j\big) \;\ge\; 1 - \cos\big(2\pi/W_{\max}\big),$$

the second inequality because $W_j \le W_{\max}$ makes the angle $2\pi/W_j$ no smaller than $2\pi/W_{\max}$, both lying in $(0, \pi]$ whenever $W_j \ge 2$. Hence $\varepsilon \ge 1 - \cos(2\pi/W_{\max})$. □

*Example.* Lemma E.3 provides a worst-case upper bound. Numerical optimization yields much tighter results for standard neighborhoods. For example, on $\mathbb{Z}_6^2$, the von Neumann, Moore, and eight-offset routing sets $\mathcal{U}$ can each be isolated using only four characters $\{\pm e_1, \pm e_2\}$, significantly below the guaranteed 32 characters. This setting is analyzed thoroughly in §E.5.

If the bounding box strictly contains the target set, the box kernel constitutes the first stage of a two-stage probabilistic construction.

**Theorem E.4** (Grid-independent spatial isolation)**.** *Let $\emptyset \ne \mathcal{N} \subsetneq \mathcal{G}$ and $\gamma > 0$. There is a $g : \mathcal{G} \to \mathbb{R}$ isolating $\mathcal{N}$ with margin $\gamma$ and*

$$|\operatorname{supp}\hat{g} \setminus \{0\}| \;\le\; 2d\,|\mathcal{N}| \;+\; 2\nu, \qquad \nu \;:=\; \big\lceil 40\,|\mathcal{N}|\,(d \ln|\mathcal{N}| + 1)\big\rceil,$$

*that is $O\big(|\mathcal{N}|\, d \log|\mathcal{N}|\big)$ frequencies. The count is independent of $T$ and $|\mathcal{G}|$.*

*Proof. Stage 1: a box around the target set.* Let $Q$ and $B$ be the box kernel and box of Lemma E.3. Then $Q \ge 0$ vanishes exactly on $B$, we have $\mathcal{N} \subseteq B$ and $|B| \le |\mathcal{N}|^d$, and $Q$ has at most $2d|\mathcal{N}|$ nontrivial characters. It remains to separate $\mathcal{N}$ from $B \setminus \mathcal{N}$, a set of at most $|\mathcal{N}|^d$ points; this is the only step at which the size of the grid could have entered, and Stage 3 is where it is kept out.

*Stage 2: one random character.* Fix $x \in B$, let $\omega$ be uniform on $\widehat{\mathcal{G}}$, and put

$$Z(\omega) \;:=\; \sum_{u \in \mathcal{N}} \cos\varphi_\omega(x - u) \;=\; \Re \sum_{u \in \mathcal{N}} \chi_\omega(x - u). \tag{17}$$

By (11) the term indexed by $u$ has expectation 1 if $x = u$ and 0 otherwise, so

$$\mathbb{E}Z \;=\; |\{u \in \mathcal{N} : u = x\}| \;=\; \begin{cases} 1 & x \in \mathcal{N}, \\ 0 & x \in B \setminus \mathcal{N}. \end{cases} \tag{18}$$

The two cases differ by 1, and the rest of the proof turns that gap into a deterministic one.

For the second moment, apply the product-to-sum identity $\cos a \cos b = \frac{1}{2}[\cos(a-b) + \cos(a+b)]$ to each pair of terms of (17) and use $\varphi_\omega(x-u) \pm \varphi_\omega(x-v) = \varphi_\omega\big((x-u) \pm (x-v)\big)$ from (10):

$$\mathbb{E}[Z^2] \;=\; \tfrac{1}{2} \sum_{u,v \in \mathcal{N}} \mathbb{E}\cos\varphi_\omega(v-u) \;+\; \tfrac{1}{2} \sum_{u,v \in \mathcal{N}} \mathbb{E}\cos\varphi_\omega(2x-u-v). \tag{19}$$

Each expectation in (19) is 1 or 0, again by (11). In the first sum the surviving pairs are those with $v = u$, of which there are $|\mathcal{N}|$. In the second the surviving pairs are those with $u + v = 2x$; here $u$ determines $v$, so at most $|\mathcal{N}|$ pairs survive. Hence

$$\operatorname{Var} Z \;\le\; \mathbb{E}[Z^2] \;\le\; \tfrac{1}{2}|\mathcal{N}| + \tfrac{1}{2}|\mathcal{N}| \;=\; |\mathcal{N}|. \tag{20}$$

Finally $|Z| \le |\mathcal{N}|$ termwise, so with (18),

$$|Z - \mathbb{E}Z| \;\le\; |\mathcal{N}| + 1 \;\le\; 2|\mathcal{N}|. \tag{21}$$

*Stage 3: sampling, and why the grid does not enter.* Draw $\omega^{(1)}, \dots, \omega^{(\nu)}$ independent and uniform on $\widehat{\mathcal{G}}$ and put $R \coloneqq \nu^{-1} \sum_{i \le \nu} Z(\omega^{(i)})$. Bernstein's inequality for independent, mean-zero summands with variance proxy $\sigma^2$ and range bound $M$ reads

$$\mathbb{P}\big(|R - \mathbb{E}R| \ge t\big) \;\le\; 2\exp\!\left(-\frac{\nu t^2}{2\big(\sigma^2 + Mt/3\big)}\right).$$

Substituting $\sigma^2 = |\mathcal{N}|$ from (20), $M = 2|\mathcal{N}|$ from (21) and $t = \frac{1}{4}$,

$$\begin{aligned}\mathbb{P}\big(|R - \mathbb{E}R| \ge \tfrac{1}{4}\big) \;&\le\; 2\exp\!\left(-\frac{\nu/16}{2\big(|\mathcal{N}| + |\mathcal{N}|/6\big)}\right) \;=\; 2\exp\!\left(-\frac{\nu/16}{\frac{7}{3}|\mathcal{N}|}\right) \\ &=\; 2\exp\!\left(-\frac{3\nu}{112\,|\mathcal{N}|}\right).\end{aligned} \tag{22}$$

This bound holds for each fixed $x \in B$. Taking a union bound over the at most $|\mathcal{N}|^d$ points of $B$, the probability that $|R - \mathbb{E}R| \ge \frac{1}{4}$ somewhere on $B$ is less than 1 as soon as

$$2\,|\mathcal{N}|^d \exp\!\left(-\frac{3\nu}{112|\mathcal{N}|}\right) < 1 \qquad\Longleftrightarrow\qquad \nu \;>\; \tfrac{112}{3}\,|\mathcal{N}|\big(d\ln|\mathcal{N}| + \ln 2\big).$$

The $\nu$ of the statement satisfies this, since $40 > 112/3 = 37.33\ldots$ and $1 > \ln 2$. The grid enters neither (22), whose constants come from (20) and (21) alone, nor the union bound, which ranges over $B$ alone; the final count is therefore independent of $W_1, \dots, W_d$.

Some realization of $\omega^{(1)}, \dots, \omega^{(\nu)}$ therefore has

$$R \ge \tfrac{3}{4} \text{ on } \mathcal{N}, \qquad R \le \tfrac{1}{4} \text{ on } B \setminus \mathcal{N}, \tag{23}$$

by (18); fix it. Being an average of the real functions (17), $R$ is real, and $\operatorname{supp} \hat{R} \setminus \{0\}$ has at most $2\nu$ elements: the sampled frequencies and their conjugates. Termwise, $|R| \le |\mathcal{N}|$ everywhere on $\mathcal{G}$.

*Stage 4: synthesis.* Let $\varepsilon \coloneqq \min_{x \notin B} Q(x) > 0$, a minimum over a finite set on which $Q$ does not vanish, and put

$$g_0 \;\coloneqq\; R \;-\; \frac{|\mathcal{N}|}{\varepsilon}\, Q \;-\; \tfrac{1}{2}. \tag{24}$$

The three regions of $\mathcal{G}$ are handled in turn.

(a) $x \in \mathcal{N}$. Then $Q(x) = 0$ because $\mathcal{N} \subseteq B$, so $g_0(x) = R(x) - \frac{1}{2} \ge \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$ by (23).

(b) $x \in B \setminus \mathcal{N}$. Again $Q(x) = 0$, so $g_0(x) = R(x) - \frac{1}{2} \le \frac{1}{4} - \frac{1}{2} = -\frac{1}{4}$.

(c) $x \notin B$. Then $Q(x) \ge \varepsilon$, so $(|\mathcal{N}|/\varepsilon)Q(x) \ge |\mathcal{N}|$, while $R(x) \le |\mathcal{N}|$; hence $g_0(x) \le |\mathcal{N}| - |\mathcal{N}| - \frac{1}{2} = -\frac{1}{2} < -\frac{1}{4}$.

So $g_0 \ge \frac{1}{4}$ on $\mathcal{N}$ and $g_0 \le -\frac{1}{4}$ off it, and $g \coloneqq 4\gamma\, g_0$ isolates $\mathcal{N}$ with margin $4\gamma(\frac{1}{4} + \frac{1}{4}) = 2\gamma \ge \gamma$. Its spectrum is contained in $\operatorname{supp} \hat{R} \cup \operatorname{supp} \hat{Q} \cup \{0\}$, so $|\operatorname{supp} \hat{g} \setminus \{0\}| \le 2\nu + 2d|\mathcal{N}|$, which is the stated count.

Two degenerate cases are worth recording. If $B = \mathcal{G}$, case (c) is vacuous, no $x$ lies outside $B$, and any $\varepsilon > 0$ serves in (24). If $B = \mathcal{N}$, case (b) is vacuous and the random part is unnecessary altogether: Lemma E.3(i) already isolates $\mathcal{N}$, at $2d|\mathcal{N}|$ frequencies and with no $\nu$ term. □

*Remark.* Because spectral support cannot exceed $|\mathcal{G}|$, Theorem E.4 is informative when $|\mathcal{N}|\, d \log|\mathcal{N}| \ll |\mathcal{G}|$, the regime where grid independence is critical. The theorem establishes a grid-independent scaling law irrespective of offset topology.

### E.3 Attention concentration from a score gap

While isolation strictly orders target offsets above others, securing a quantitative guarantee on attention mass requires scaling. This is the only point in the analysis where the absolute scale of the positional kernel matters; previous results depend solely on the Fourier spectral support.

**Corollary E.5** (Softmax concentration)**.** *Let a softmax be taken over a finite set of keys scored $\Lambda g$, where $\Lambda > 0$ and $g$ isolates a target subset of the keys with margin $\gamma$. Suppose at least one target key is present and at most $N_{\neq}$ non-target keys are. Then the attention mass placed off the target subset is at most $N_{\neq} e^{-\Lambda\gamma}$, and is therefore at most $\varepsilon$ as soon as*

$$\Lambda\gamma \;\geq\; \log(N_{\neq}/\varepsilon), \qquad \textit{equivalently} \qquad \Lambda \;\geq\; \gamma^{-1}\log(N_{\neq}/\varepsilon). \tag{25}$$

*Proof.* Write $s_\star \coloneqq \min_{\text{target}} g$ and $s_{\neq} \coloneqq \max_{\text{non-target}} g$, so that isolation reads $s_\star - s_{\neq} \geq \gamma$ and, after scaling, $\Lambda s_\star - \Lambda s_{\neq} \geq \Lambda\gamma$. Split the partition function into its target and non-target parts,

$$A \;\coloneqq \sum_{\text{target keys}} e^{\Lambda g}, \qquad B \;\coloneqq \sum_{\text{non-target keys}} e^{\Lambda g}.$$

At least one target key is present, so $A \geq e^{\Lambda s_\star}$. Every non-target key contributes at most $e^{\Lambda s_{\neq}}$, and there are at most $N_{\neq}$ of them, so

$$\begin{aligned} B \;&\leq\; N_{\neq} e^{\Lambda s_{\neq}} \;=\; N_{\neq} e^{\Lambda s_\star} e^{-\Lambda(s_\star - s_{\neq})} \\ &\leq\; N_{\neq} e^{-\Lambda\gamma}\, e^{\Lambda s_\star} \;\leq\; N_{\neq} e^{-\Lambda\gamma} A. \end{aligned}$$

The mass off the target subset is $B/(A+B)$, and since $A, B \geq 0$,

$$\frac{B}{A+B} \;\leq\; \frac{B}{A} \;\leq\; N_{\neq} e^{-\Lambda\gamma}.$$

This is at most $\varepsilon$ exactly when $e^{\Lambda\gamma} \geq N_{\neq}/\varepsilon$, which is (25). □

### E.4 Selecting spatial offsets in the preceding time step

This section details the spatial positional block $B_2$ (Definition G.2), establishing the interface for Appendix G. Lemma E.1 connects spectral counts to table width. A translation-invariant positional bias yields a score matrix $P^\top W_Q^{(1)\top} W_K^{(1)} P$ (where $P$ is the positional table). Setting $Q = W_Q^{(1)\top} W_K^{(1)}$ confirms the table requires exactly $|\operatorname{supp} \hat{g}|$ coordinates. Discarding the trivial character, this is the addressing dimension $d_S$. An attention head selects its routed offset by rotating the $2 \times 2$ blocks in (15).

**Corollary E.6** (Spatiotemporal addressing interface)**.** *Let $\emptyset \neq \mathcal{N} \subseteq \mathcal{G}$ be a set of offsets and let a Layer-1 head be required to score, from its query's own cell $\kappa(n)$, the cells at $\kappa(n) + \delta$ for $\delta \in \mathcal{N}$, in the time step before the target cell's, strictly above every other causally visible cell key. There is a positional table $p_n \in \mathbb{R}^{d_p}$, indexed by sequence position, and projections $W_Q, W_K$ whose pre-softmax score does so with a margin $\gamma > 0$, where*

$$d_p \;=\; d_S + 2, \qquad d_S \;\leq\; 2d|\mathcal{N}| + 2\nu \;=\; O\big(|\mathcal{N}|\, d \log|\mathcal{N}|\big),$$

*and $d_S = 2d$ when $\mathcal{N}$ is a single offset. The width is independent of the volume and of the horizon. The two temporal coordinates carry a scale $\lambda$, which must satisfy*

$$\begin{aligned} \lambda \;&\geq\; \frac{R+\gamma}{\gamma_T}, \\ R \;\coloneqq\; \max_{\delta\in\mathcal{G}} g(\delta) - \min_{\delta\in\hat{u}_d - \mathcal{N}} g(\delta) \;&\geq\; 0, \qquad \gamma_T \;\coloneqq\; 1 - \cos(2\pi/T), \end{aligned} \tag{26}$$

*where $g$ is the spatial kernel. The scale $\lambda$ enters no width.*

*Proof. Stage 1: the group and the target set.* A query at sequence position $n$ has virtual time step $\tau(n)$ and virtual cell $\kappa(n)$, both defined in Appendix C.1, and its target cell is $\kappa(n) + \hat{u}_d$ (Lemma C.2). It must score the keys $n'$ with

$$\kappa(n') - \kappa(n) \in \mathcal{N} \qquad \text{and} \qquad \tau(n) - \tau(n') = 1$$

above all others. Stage 5 realizes the spatial score in the *target-minus-key* variable

$$\delta \;\coloneqq\; \kappa(n) + \hat{u}_d - \kappa(n'),$$

and a key at $\kappa(n') = \kappa(n) + \delta'$ has $\delta = \hat{u}_d - \delta'$, so the first condition reads $\delta \in \hat{u}_d - \mathcal{N}$. Above, $\hat{u}_d - \mathcal{N} \coloneqq \{\hat{u}_d - \delta' : \delta' \in \mathcal{N}\}$, and $\delta' \mapsto \hat{u}_d - \delta'$ is a bijection of $\mathcal{G}$, so $|\hat{u}_d - \mathcal{N}| = |\mathcal{N}|$ and every width bound of the statement, which is stated in $|\mathcal{N}|$, is unaffected. Each condition constrains a difference, so the ambient group is $\mathcal{G} \times \mathbb{Z}_T$ and the target set is $(\hat{u}_d - \mathcal{N}) \times \{1\}$. We build a kernel on that product, one block per factor.

*Stage 2: the spatial block.* Let $g$ isolate $\hat{u}_d - \mathcal{N}$ in $\mathcal{G}$ with margin $\gamma$, by Theorem E.4, and let $\Omega \coloneqq \operatorname{supp} \hat{g} \setminus \{0\}$ with $|\Omega| = d_S$. Removing the trivial character replaces $g$ by the centered kernel $g - \hat{g}(0)$, which isolates the same set with the same margin (Definition E.2); we write $g$ for it from here on. Let $\Omega_+ \subseteq \Omega$ contain one representative of each orbit $\{\omega, -\omega\}$. Write $g$ in the real form (13) with coefficients $\alpha_\omega, \beta_\omega$, and take the coordinates and blocks Lemma E.1 supplies, indexed by the representatives:

$$\phi(z) \;\coloneqq\; \big(e_\omega(z)\big)_{\omega\in\Omega_+} \;=\; \Big(\cos\varphi_\omega(z),\ \sin\varphi_\omega(z)\Big)_{\omega\in\Omega_+} \;\in\; \mathbb{R}^{d_S}, \qquad M_\omega \;=\; \begin{bmatrix} \alpha_\omega & -\beta_\omega \\ \beta_\omega & \alpha_\omega \end{bmatrix}, \tag{27}$$

with the sine coordinate omitted for a self-paired representative. This gives one $2 \times 2$ block per conjugate pair and one $1 \times 1$ block per self-paired character, for $|\Omega| = d_S$ coordinates in total. By (16), $\phi(z)^\top \operatorname{blkdiag}(M_\omega)_\omega\, \phi(z') = g(z - z')$.

*Stage 3: the temporal block.* Step selection is isolation of the single displacement 1 in $\mathbb{Z}_T$, which Lemma E.3(ii) covers with $d = 1$, $W_1 = T$ and $\mathcal{N} = \{1\}$. Discarding the constant term, which Definition E.2 permits, the kernel is

$$h(s) \;\coloneqq\; \cos\big(2\pi(s-1)/T\big), \qquad h(1) = 1, \qquad h(s) \le \cos(2\pi/T) = 1 - \gamma_T \;\; (s \neq 1), \tag{28}$$

so $h$ isolates $\{1\}$ with margin $\gamma_T = 1 - \cos(2\pi/T) > 0$, whatever $T \ge 2$ is. Its spectrum is $\{\pm 1\} \subseteq \widehat{\mathbb{Z}_T}$, at most two characters (one when $T = 2$, where $1 = -1$), so Lemma E.1 realizes it in at most two coordinates. We always use two, padding with the identically zero sine coordinate when $T = 2$:

$$\psi(\tau) \;\coloneqq\; \begin{bmatrix} \cos(2\pi\tau/T) \\ \sin(2\pi\tau/T) \end{bmatrix} \in \mathbb{R}^2, \qquad M_{\text{time}} \;\coloneqq\; \begin{bmatrix} \cos(2\pi/T) & -\sin(2\pi/T) \\ \sin(2\pi/T) & \cos(2\pi/T) \end{bmatrix} \;=\; R(2\pi/T), \tag{29}$$

the rotation by one time step: reading (16) with $\alpha_\omega = \cos(2\pi/T)$ and $\beta_\omega = \sin(2\pi/T)$ gives $\psi(\tau)^\top M_{\text{time}} \psi(\tau') = \cos\big(2\pi(\tau - \tau' - 1)/T\big) = h(\tau - \tau')$.

*Stage 4: the table and the bilinear form.* Concatenate the two blocks. The table entry at sequence position $n$ is

$$p_n \;\coloneqq\; \begin{bmatrix} \psi(\tau(n)) \\ \phi(\kappa(n)) \end{bmatrix} \;\in\; \mathbb{R}^{d_S+2}, \qquad W_Q^\top W_K \;\coloneqq\; \operatorname{blkdiag}\big(\lambda M_{\text{time}},\, (M_\omega)_{\omega\in\Omega_+}\big), \tag{30}$$

evaluated at the virtual time step and virtual cell of the position, which Appendix C.1 defines for cell tokens and separators alike. Because $W_Q^\top W_K$ is block diagonal with the blocks aligned to the blocks of $p_n$, the quadratic form is the sum of the per-block forms of Stages 2 and 3:

$$\begin{aligned} p_n^\top \big(W_Q^\top W_K\big) p_{n'} \;&=\; \lambda\, \psi(\tau(n))^\top M_{\text{time}}\, \psi(\tau(n')) \;+\; \sum_{\omega\in\Omega_+} e_\omega(\kappa(n))^\top M_\omega\, e_\omega(\kappa(n')) \\ &=\; g\big(\kappa(n) - \kappa(n')\big) \;+\; \lambda\, h\big(\tau(n) - \tau(n')\big). \end{aligned} \tag{31}$$

This is the form Appendix C.2 quotes as $p_n = (\phi(\kappa(n)), \psi(\tau(n)))$.

*Stage 5: the successor shift costs no coordinate.* The spatial argument used in Stage 1 is $\kappa(n) + \hat{u}_d - \kappa(n')$, whereas (31) delivers $g(\kappa(n) - \kappa(n'))$. The discrepancy is the constant $\hat{u}_d$, and by (10) a constant displacement is a fixed rotation of every character:

$$\chi_\omega(z + \hat{u}_d) \;=\; \chi_\omega(\hat{u}_d)\, \chi_\omega(z), \qquad \text{that is} \qquad \varphi_\omega(z + \hat{u}_d) = \varphi_\omega(z) + \theta_\omega, \quad \theta_\omega \coloneqq \varphi_\omega(\hat{u}_d).$$

The product $M_\omega R(-\theta_\omega)$ is again a scaled rotation, by the last paragraph of the proof of Lemma E.1. Replacing each block $M_\omega$ of (30) by it turns (16) into $\alpha_\omega \cos(A - B + \theta_\omega) + \beta_\omega \sin(A - B + \theta_\omega)$,

so the spatial term of (31) becomes $g(\kappa(n) + \hat{u}_d - \kappa(n'))$ as required. The table itself is unchanged and no coordinate is added. By Lemma C.2 the offset is the same constant $\hat{u}_d$ for every position which has a target cell, separators included, so one choice of blocks serves all of them; the separator case is recorded after this proof.

*Stage 6: the margin.* Write the score of key $n'$ at query $n$ as $\Lambda\, G(\delta, s)$ with $\delta \coloneqq \kappa(n) + \hat{u}_d - \kappa(n')$, $s \coloneqq \tau(n) - \tau(n')$ and

$$G(\delta, s) \;\coloneqq\; g(\delta) \;+\; \lambda\, h(s),$$

which Stages 4 and 5 realize. Put $m \coloneqq \min_{\delta \in \hat{u}_d - \mathcal{N}} g(\delta)$, so that $\max_{\delta \notin \hat{u}_d - \mathcal{N}} g \le m - \gamma$ by isolation and $\max_{\delta \in \mathcal{G}} g = m + R$ by the definition of $R$ in (26). Three cases exhaust the keys.

(a) *Target keys*, $\delta \in \hat{u}_d - \mathcal{N}$ and $s = 1$. Then $h(s) = 1$ and $G \ge m + \lambda$.

(b) *Right step, wrong offset*, $\delta \notin \hat{u}_d - \mathcal{N}$ and $s = 1$. Then $G \le (m - \gamma) + \lambda$, which is below case (a) by $\gamma$.

(c) *Wrong step*, $s \neq 1$. Then $h(s) \le 1 - \gamma_T$ by (28) and $g(\delta) \le m + R$, so

$$G \;\le\; m + R + \lambda(1 - \gamma_T) \;=\; m + \lambda + \big(R - \lambda\gamma_T\big) \;\le\; m + \lambda - \gamma,$$

the last step by the choice (26) of $\lambda$, which is exactly $\lambda\gamma_T \ge R + \gamma$. Again this is below case (a) by $\gamma$.

So $G$ isolates $(\hat{u}_d - \mathcal{N}) \times \{1\}$ with margin $\gamma$, and the pre-softmax score $\Lambda G$ isolates it with margin $\Lambda\gamma$. Case (c) is the one which forces the scaling: with $\lambda = 1$ the bound there is $m + 1 + R - \gamma_T$, which exceeds the target floor $m + 1$ whenever $R > \gamma_T$, that is, whenever the spatial kernel varies across $\hat{u}_d - \mathcal{N}$ by more than one time step is worth. The causal mask is not needed for any of this, since $h$ orders the whole of $\mathbb{Z}_T$, visible part included.

*Widths.* The two blocks of (30) occupy orthogonal coordinates, so their ranks add and $d_p = d_S + 2$. The count $d_S \le 2d|\mathcal{N}| + 2\nu$ is Theorem E.4, and $d_S = 2d$ for a single offset is Lemma E.3(ii) on $\mathcal{G}$. Neither count involves the volume or the horizon: the grid is absent from Theorem E.4 by Stage 3 of its proof, and the horizon is absent because the temporal block has two coordinates for every $T$. The scale $\lambda$ of (26) multiplies a block; it does not widen one. □

*Remark* E.7 (Separator keys). The guarantee above is stated over cell keys. A separator carries the virtual address of the cell one step before its successor (Lemma C.4*(i)*), so a separator key can hold the same $(\tau, \kappa)$ as a target cell, and a score built from $\tau$ and $\kappa$ alone cannot order it below. Proposition G.6 adds a rank-one gate of strength $\Gamma_1$ which places every separator key $\Lambda\gamma$ below the target floor, extending the corollary to all causally visible keys at no positional width.

*Remark* (Boundary cases of the interface). We clarify three assumptions used above: *Horizon.* Stage 3 requires $T \ge 2$ so that $\gamma_T = 1 - \cos(2\pi/T) > 0$, ensuring the temporal kernel has a nonempty complement. Definition 1 imposes $T \ge 2$; at $T = 1$, the prediction task is empty. *Full target set.* While Definition E.2 and Theorem E.4 require $\mathcal{N} \subsetneq \mathcal{G}$, the case $\mathcal{N} = \mathcal{G}$ is covered by setting $g \equiv 0$ (so $d_S = 0$ and $d_p = 2$). The temporal block alone correctly identifies all keys in the previous time step with margin $\gamma = \lambda\gamma_T$, making the interface total in $\mathcal{N}$. *Degenerate axes.* If $W_j = 2$, Lemma E.3(ii) contributes one character instead of two on that axis, giving $d_S \le 2d$ without increasing any stated width.

On the query side, the positional table accommodates separator positions. By Lemma C.4, a separator at $n$ carries virtual indices $\kappa(n)$ and $\tau(n)$, targets $\kappa(n) + \hat{u}_d$, and causally accesses all parents in step $\tau(n) - 1$. Because the proof depends on the query solely through $\kappa$ and $\tau$, (30) applies verbatim to separator queries without additional coordinates, maintaining the margin $\gamma$. Remark E.7 covers the key side.

Subsequent sections depend on the spatial kernel $g$, its isolation margin $\gamma$, the positional table width $d_p = d_S + 2$, and the separator gate of Proposition G.6. Corollary E.5 converts this margin and logit scale into an attention mass bound. A caller simply specifies the desired offsets $\mathcal{N}$ relative to the query; Stage 1 of the proof handles the translation to the kernel's variable.

### E.5 Example: addressing the 2D von Neumann routing set

The preceding bounds are worst-case over all subsets $\mathcal{N}$. Here, we instantiate the derivation for a specific experimental setting with explicit numerical constants.

**Example E.8** (The routing set on $\mathbb{Z}_6^2$)**.** *Let $\mathcal{G} = \mathbb{Z}_6^2$, let $D_K$ be the von Neumann neighborhood, let $\hat{u}_d = (0,1)$ and let $T = 12$, so that the routing set $\mathcal{U} = D_K \cup (D_K + \hat{u}_d)$ of Definition 2 has the $|\mathcal{U}| = 8$ elements*

$$\mathcal{U} \;=\; \{(0,0),\,(0,1),\,(0,2),\,(0,5),\,(1,0),\,(1,1),\,(5,0),\,(5,1)\}.$$

*Then the anisotropic, half-shifted degree-one kernel*

$$g(\delta) \;=\; \tfrac{3}{2}\cos\big(2\pi\delta_1/6\big) \;+\; \cos\big(2\pi(\delta_2 - \tfrac{1}{2})/6\big) \tag{32}$$

*isolates $\mathcal{U}$ with margin $\gamma = \frac{3}{4}$ at the four characters $\Omega = \{\pm e_1, \pm e_2\}$. Here $\hat{u}_d - \mathcal{U} = \mathcal{U}$, so this is also the set Stage 2 of Corollary E.6 asks $g$ to isolate when the corollary is invoked with $\mathcal{N} = \mathcal{U}$. Thus*

$$d_S \;=\; 4, \qquad d_p \;=\; 6,$$

*and Corollary E.6 holds at these widths with temporal scale $\lambda = 13$. At the stream length $N = 504$ of this setting, a logit scale $\Lambda = 12$ places mass at least $1 - \frac{1}{12}$ on $\mathcal{U}$.*

*Proof. The spectrum and the blocks.* Expanding the half-step shift by the angle-difference identity, with $2\pi \cdot \frac{1}{2}/6 = \pi/6$,

$$\begin{aligned}\cos\big(2\pi(\delta_2 - \tfrac{1}{2})/6\big) \;&=\; \cos\big(\varphi_{e_2}(\delta) - \tfrac{\pi}{6}\big)\\ &=\; \cos\tfrac{\pi}{6}\,\cos\varphi_{e_2}(\delta) \;+\; \sin\tfrac{\pi}{6}\,\sin\varphi_{e_2}(\delta),\end{aligned}$$

so (32) is in the real form (13) with

$$\alpha_{e_1} = \tfrac{3}{2}, \quad \beta_{e_1} = 0, \qquad \alpha_{e_2} = \tfrac{\sqrt{3}}{2}, \quad \beta_{e_2} = \tfrac{1}{2}.$$

Both coefficients are nonzero, so $\operatorname{supp}\hat{g} \setminus \{0\} = \{\pm e_1, \pm e_2\}$ and $d_S = |\Omega| = 4$: one conjugate pair per axis, hence the two $2 \times 2$ blocks of (15),

$$M_{e_1} \;=\; \begin{bmatrix} \frac{3}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix} \;=\; \tfrac{3}{2}\,I_2, \qquad M_{e_2} \;=\; \begin{bmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \;=\; R(\pi/6).$$

The shift costs no coordinate: it is absorbed into the rotation angle of the second block, as Lemma E.1 shows in general. With the two temporal coordinates of Stage 3 of Corollary E.6, $d_p = d_S + 2 = 6$.

*The margin.* On $\mathbb{Z}_6$ the cosine takes the four values $1, \frac{1}{2}, -\frac{1}{2}, -1$, so $g$ takes finitely many values which can be listed. On $\mathcal{U}$ it takes three:

$$g(0,0) = g(0,1) = \tfrac{3}{2} + \tfrac{\sqrt{3}}{2}, \qquad g(0,2) = g(0,5) = \tfrac{3}{2},$$

$$g(1,0) = g(1,1) = g(5,0) = g(5,1) = \tfrac{3}{4} + \tfrac{\sqrt{3}}{2},$$

so $\min_{\mathcal{U}} g = \frac{3}{2}$. Off $\mathcal{U}$ the largest value is $\max_{\delta \notin \mathcal{U}} g = \frac{3}{4}$, attained at the four offsets $(\pm 1, 2)$ and $(\pm 1, 5)$. Hence

$$\gamma \;=\; \min_{\mathcal{U}} g - \max_{\notin \mathcal{U}} g \;=\; \tfrac{3}{2} - \tfrac{3}{4} \;=\; \tfrac{3}{4}.$$

The half-step shift makes this work. $\mathcal{U}$ is asymmetric along the innermost axis, containing $(0,2)$ but not $(0,4)$, and an unshifted kernel, being even in $\delta_2$, scores those two offsets alike and cannot separate them.

*The temporal scale.* The largest value of $g$ on the whole grid is $\max_{\mathcal{G}} g = \frac{3}{2} + \frac{\sqrt{3}}{2}$, attained at $(0,0) \in \mathcal{U}$, so

$$R \;=\; \max_{\mathcal{G}} g - \min_{\mathcal{U}} g \;=\; \tfrac{\sqrt{3}}{2} \;\approx\; 0.866, \qquad \gamma_T \;=\; 1 - \cos(2\pi/12) \;=\; 1 - \tfrac{\sqrt{3}}{2} \;\approx\; 0.134,$$

and (26) requires

$$\lambda \;\geq\; \frac{R + \gamma}{\gamma_T} \;=\; \frac{\frac{\sqrt{3}}{2} + \frac{3}{4}}{1 - \frac{\sqrt{3}}{2}} \;\approx\; 12.06,$$

so $\lambda = 13$ suffices. The gap between $\gamma = \frac{3}{4}$ and $\gamma_T \approx 0.134$ makes the scaling necessary here: an unscaled sum of the two kernels would score a key in the wrong step at up to $\frac{3}{2} + \frac{\sqrt{3}}{2} + \cos(\pi/6) \approx 3.23$, above the target floor $\frac{3}{2} + 1 = \frac{5}{2}$.

*The logit scale.* The stream length for this setting is $N = T(|\mathcal{G}| + |\mathcal{G}|/W_2) = 12\,(36 + 6) = 504$ (Definition C.1). Corollary E.5 with $N_{\neq} \le N$ and off-target mass $\varepsilon = \frac{1}{12}$ gives

$$\Lambda \;\ge\; \gamma^{-1} \log(N/\varepsilon) \;=\; \tfrac{4}{3} \log(6048) \;\approx\; 11.61,$$

so $\Lambda = 12$ places all but $\frac{1}{12}$ of the attention mass on $\mathcal{U}$. We chose $\varepsilon = \frac{1}{12}$ to make the arithmetic concrete. This $\mathcal{N} = \mathcal{U}$ addressing serves Construction II, which decodes each block exactly and so requires (36) of $\Lambda$ instead. At $V = 2$, $M = 12$ colors and $n_{\max} \le T\lceil |\mathcal{G}|/M \rceil = 36$, that condition reads

$$\Lambda\gamma \;\ge\; \log\bigl(2V(n_{\max} - 1)\bigr) \;=\; \log 140 \;\approx\; 4.94,$$

so $\Lambda \ge 6.59$ and the same $\Lambda = 12$ satisfies it. □

*Remark.* Two comparisons relate this example to the general statements.

First, compared to Lemma E.3: the same four characters isolate the von Neumann and Moore neighborhoods on this grid (margin $\frac{1}{2}$) via the isotropic unshifted kernel $g(\delta) = \cos\varphi_{e_1}(\delta) + \cos\varphi_{e_2}(\delta)$ ($M_{e_1} = M_{e_2} = I_2$).

Second, the theorem becomes informative only when $|\mathcal{N}|d\log|\mathcal{N}|$ is much less than the volume, the regime where grid independence is critical. For the small grids in Table 4, explicit numerical search provides tighter bounds. In 1D, no search is required: the bounding box equals the set itself, so Lemma E.3(i) applies directly, avoiding Theorem E.4's probabilistic bounds, and yielding $d_S \le 2d|\mathcal{U}| = 8$.

# F Colorings for Shared-head decoding

Construction II (Appendix G) employs a shared Layer-1 head. A key at cell $z'$ projects a one-hot *color code* $\varphi(z') = e_{c(z')} \in \mathbb{R}^M$ via a map $c : \mathcal{G} \to [M]$. The head's output at query $z$ is a vector $a \in \mathbb{R}^{M+1}$ (comprising $M$ color coordinates and one reference coordinate) which superposes all $|\mathcal{U}|$ gathered cells. The cell at offset $\delta$ writes to coordinate $c(z + \delta)$ with a state-dependent weight. For the MLP to recover this cell, no other gathered cell may occupy the same coordinate, requiring:

$$\delta \;\longmapsto\; c(z + \delta) \qquad \text{is injective on } \mathcal{U}, \text{ for every } z \in \mathcal{G}.$$

This condition defines an admissible coloring (Definition F.1). This section characterizes these maps and bounds the required number of colors $M$ as a function of $\mathcal{G}$ and $\mathcal{U}$.

A cell's color is a function of its position, so the key and the query each read it from the positional table. The key reads the color of its own cell, and the query reads the color it expects at each routed offset. The two uses occupy separate coordinates, which Corollary F.5 counts in parts (e) and (b). Appendix G assigns those coordinates to residual blocks and wires them into the value projection and the MLP.

## F.1 Admissible colorings and the Cayley graph

**Definition F.1** (Admissible coloring)**.** A *coloring* is a map $c : \mathcal{G} \to [M]$, and it is *admissible for* $\mathcal{U}$ if it is injective on every translate $z + \mathcal{U}$. Equivalently, $c$ is a proper coloring of the Cayley graph

$$\Gamma_{\mathcal{U}} \;\coloneqq\; \mathrm{Cay}\bigl(\mathcal{G},\; (\mathcal{U} - \mathcal{U}) \setminus \{0\}\bigr),$$

whose vertices are the cells and whose edges join cells differing by a nonzero element of $\mathcal{U} - \mathcal{U}$. We write $M^\star$ for the least $M$ admitting an admissible coloring, that is, for the chromatic number of $\Gamma_{\mathcal{U}}$. A coloring is *periodic* if it is the quotient map $\chi : \mathcal{G} \to \mathcal{G}/K$ by a subgroup $K \le \mathcal{G}$, composed with any bijection $\mathcal{G}/K \cong [M_K]$, equivalently if its color classes are exactly the cosets of $K$, and its number of colors is then $M_K \coloneqq |\mathcal{G}/K|$. The choice of bijection is immaterial: relabeling preserves both admissibility and the number of colors.

Table 6: The coloring Construction II uses in each experimental setting. Every one is admissible by Lemma F.2, and every one is far below the degree bound on the same line.

| | 1D, $V{=}2$ | 1D, $V{=}3$ | 2D von Neumann |
|---|---|---|---|
| $\lvert(\mathcal{U}{-}\mathcal{U}) \setminus \{0\}\rvert$ | 6 | 6 | 21 |
| degree bound on $M^\star$ | 7 | 7 | 22 |
| subgroup $K$ | $4\mathbb{Z}_{16}$ | $4\mathbb{Z}_{32}$ | $\langle(2,2)\rangle$ |
| colors $M = M_K$ | 4 | 4 | 12 |

Because a query gathers uniquely labeled cells, two cells may share a label only if their nonzero difference lies outside $\mathcal{U} - \mathcal{U}$, reducing the problem to graph coloring.

Periodic colorings are easily constructed, and their admissibility reduces to an algebraic condition:

**Lemma F.2** (When a periodic coloring is admissible)**.** *A periodic coloring $\chi : \mathcal{G} \to \mathcal{G}/K$ is admissible for $\mathcal{U}$ if and only if $K \cap (\mathcal{U} - \mathcal{U}) = \{0\}$.*

*Proof.* The equality $\chi(z + \delta) = \chi(z + \delta')$ holds exactly when $\delta - \delta' \in K$. As $\delta, \delta'$ range over $\mathcal{U}$, injectivity on every translate $z + \mathcal{U}$ fails if and only if a nonzero element of $\mathcal{U} - \mathcal{U}$ lies in $K$. □

§F.3 uses this criterion to construct admissible colorings for our experimental grids; these are significantly smaller than the worst-case bound below.

### F.2 Bounds on the number of colors

**Lemma F.3** (Bounds on the number of colors)**.**

$$|\mathcal{U}| \;\le\; M^\star \;\le\; \big|(\mathcal{U}-\mathcal{U}) \setminus \{0\}\big| + 1 \;\le\; |\mathcal{U}|^2 - |\mathcal{U}| + 1,$$

*and $M^\star \le M_K$ for every periodic admissible coloring. Moreover an admissible coloring with $|(\mathcal{U} - \mathcal{U}) \setminus \{0\}| + 1$ colors may be taken* equitable*, its color classes differing in size by at most one. The outermost bound $|\mathcal{U}|^2 - |\mathcal{U}| + 1$ does not mention the grid.*

*Remark* F.4 (Which bounds are grid-independent). The middle quantity $|(\mathcal{U} - \mathcal{U}) \setminus \{0\}|$, the degree of $\Gamma_{\mathcal{U}}$, evaluates differences *on the torus* and thus depends on side lengths (distinct differences in $\mathbb{Z}^d$ can collide modulo $W_j$, strictly decreasing the count). Evaluating this expression in $\mathbb{Z}^d$ yields a grid-independent upper bound; the figures 7, 23, and 35 below derive from this grid-independent evaluation.

For the von Neumann routing set, the $\mathbb{Z}^2$ count is $|(\mathcal{U} - \mathcal{U}) \setminus \{0\}| = 22$ (degree bound 23). On $\mathbb{Z}_6^2$, offsets $+3$ and $-3$ are identified, dropping the exact count to 21 (degree bound 22). While both bounds validly majorize $M^\star$, only the $\mathbb{Z}^d$ bound is grid-independent. Grid independence is a statement about widths. The loads and logit scales of Corollary F.5*(c)* and (36) do depend on the volume and the horizon.

*Proof.* The lower bound holds because $\mathcal{U}$ forms a clique in $\Gamma_{\mathcal{U}}$. $\Gamma_{\mathcal{U}}$ is regular of degree $\Delta = |(\mathcal{U} - \mathcal{U}) \setminus \{0\}|$; a greedy coloring uses at most $\Delta + 1$ labels. Every nonzero difference comes from one of the $|\mathcal{U}|(|\mathcal{U}| - 1)$ ordered pairs of distinct offsets, so $\Delta \le |\mathcal{U}|^2 - |\mathcal{U}|$. Equitability at $\Delta + 1$ colors follows from the Hajnal–Szemerédi theorem (Hajnal, 1970). Lemma F.2 confirms that a periodic admissible coloring is proper for $\Gamma_{\mathcal{U}}$, so $M^\star \le M_K$. □

*Remark.* The upper bound depends only on local neighborhood geometry. Evaluated in $\mathbb{Z}^d$ (Remark F.4), it requires at most 7 colors for the elementary neighborhood, 23 for von Neumann, and 35 for Moore, independent of grid side length. Theorem G.10 leverages this grid-independent bound.

### F.3 Colorings for the experimental grids

Lemma F.2 supplies admissible colorings for the three experimental settings of Appendix B.4, and each uses far fewer colors than the degree bound on the same line. Table 6 records them.

Table 7: Colorings across grids, for reference. $M_K$ is the least number of colors over periodic admissible colorings and is arithmetic in the side length. The degree bound $|(\mathcal{U}-\mathcal{U})\setminus\{0\}|+1$ of Lemma F.3 is evaluated with the differences taken in $\mathbb{Z}^d$, so it is the same for every grid and bounds $M^\star$ from above; the exact torus count can be smaller, as Remark F.4 shows on $\mathbb{Z}_6^2$. The bound never degenerates and $M_K$ sometimes does: on $\mathbb{Z}_5^2$ and $\mathbb{Z}_7^2$ every nontrivial subgroup has a nonzero intersection with $\mathcal{U}-\mathcal{U}$, so $M_K=|\mathcal{G}|$ and every cell takes a color of its own.

| neighborhood | grid | $\lvert\mathcal{U}\rvert$ | least periodic $M_K$ | degree bound | $\lvert\mathcal{G}\rvert$ |
|---|---|---|---|---|---|
| elementary | $\mathbb{Z}_5$ | 4 | 5 | 7 | 5 |
| elementary | $\mathbb{Z}_{16}$ | 4 | 4 | 7 | 16 |
| elementary | $\mathbb{Z}_{32}$ | 4 | 4 | 7 | 32 |
| von Neumann | $\mathbb{Z}_5^2$ | 8 | 25 | 23 | 25 |
| von Neumann | $\mathbb{Z}_6^2$ | 8 | 12 | 23 | 36 |
| von Neumann | $\mathbb{Z}_7^2$ | 8 | 49 | 23 | 49 |
| von Neumann | $\mathbb{Z}_8^2$ | 8 | 8 | 23 | 64 |
| Moore | $\mathbb{Z}_6^2$ | 12 | 18 | 35 | 36 |
| Moore | $\mathbb{Z}_8^2$ | 12 | 16 | 35 | 64 |
| Moore | $\mathbb{Z}_{12}^2$ | 12 | 12 | 35 | 144 |

Each coloring is the quotient by a subgroup intersecting $\mathcal{U}-\mathcal{U}$ only at 0 (Lemma F.2). For $d=1$, $\mathcal{U}=\{-1,0,1,2\}$ yields $(\mathcal{U}-\mathcal{U})\setminus\{0\}=\{\pm1,\pm2,\pm3\}$. Because this lacks nonzero multiples of 4, reduction modulo 4 is admissible on any $\mathbb{Z}_W$ where $4\mid W$, yielding an optimal $M=4$ (Lemma F.3 requires $M^\star\geq 4$). For $d=2$, coordinatewise reduction requires $M_K=18$, while the diagonal subgroup $\langle(2,2)\rangle=\{(0,0),(2,2),(4,4)\}$ avoids $\mathcal{U}-\mathcal{U}$ to achieve $M_K=12$.

The degree row in Table 6 counts $|(\mathcal{U}-\mathcal{U})\setminus\{0\}|$ on the torus, yielding 21 on $\mathbb{Z}_6^2$. Evaluating differences in $\mathbb{Z}^d$ yields the grid-independent count of 23 (Lemma F.3, Table 7). This discrepancy arises from identifying $+3$ and $-3$ on the torus (Remark F.4). Table 7 contextualizes these grids.

*Remark.* For grids in Table 7 (excluding $\mathbb{Z}_5^2$ and $\mathbb{Z}_7^2$), periodic coloring requires fewer colors and is preferred in practice. The degree bound ensures grid independence (Theorem G.10).

### F.4 Coloring interface for Shared-head decoding

The Shared-head construction relies on the following five guarantees. They fix the properties and the coordinate counts; Appendix G assigns the coordinates to residual blocks.

**Corollary F.5** (Coloring interface for the Shared-head construction)**.** *Let $c$ be an admissible coloring of $\mathcal{G}$ for $\mathcal{U}$ with $M$ colors. Then*

*(a)* Distinct slots. *For every $z$, the $|\mathcal{U}|$ cells $z+\delta$, $\delta\in\mathcal{U}$, carry pairwise distinct colors.*

*(b)* Address width. *For every $z$ the family of addresses $\big(e_{c(z+\delta)}\big)_{\delta\in\mathcal{U}}\in\mathbb{R}^{|\mathcal{U}|M}$ is the image, under one fixed affine map independent of $z$, of $|\mathcal{U}|(M-1)+1$ positional coordinates: one width-$M$ address per offset, each a one-hot and so determined by $M-1$ of its coordinates given a shared constant.*

*(c)* Load. *The number $n_{\max}$ of causally visible cells sharing one color, over a trajectory of at most $T$ time steps, satisfies*

$$n_{\max}\;\leq\;T\max_{s\in[M]}\big|c^{-1}(s)\big|.$$

*If $c$ is equitable, then $\max_s|c^{-1}(s)|\leq\lceil|\mathcal{G}|/M\rceil$, and consequently*

$$n_{\max}\;\leq\;T\left\lceil\tfrac{|\mathcal{G}|}{M}\right\rceil.$$

*(d)* Existence. *An equitable admissible $c$ exists on every grid with $M\leq|(\mathcal{U}-\mathcal{U})\setminus\{0\}|+1\leq|\mathcal{U}|^2-|\mathcal{U}|+1$, where the differences may be taken in $\mathbb{Z}^d$, so that the final bound depends on $\mathcal{U}$ and on nothing else.*

*(e)* Own-color block. *For every $z$ the own-color address $e_{c(z)}\in\mathbb{R}^M$ is a function of position alone and is carried by $M$ positional coordinates, which may be set to zero at separator positions.*

*Proof.* (a) is admissibility, which is injectivity on $z + \mathcal{U}$. For (b), the offset $\delta$ is fixed, so $z \mapsto e_{c(z+\delta)}$ depends on position alone and the family is carried by $|\mathcal{U}|$ width-$M$ positional blocks. A one-hot of width $M$ has coordinates summing to 1, so it lies in an affine subspace of dimension $M-1$ and $M-1$ coordinates determine it given the constant; the $|\mathcal{U}|$ addresses therefore take $|\mathcal{U}|(M-1)$ coordinates together with one constant coordinate.

For (c), a causally visible prefix contains at most $T$ copies of the grid, one per time step, and within a single time step the cells of color $s$ number at most $|c^{-1}(s)|$; summing over steps gives the general bound. If $c$ is equitable, the $M$ color classes of $\mathcal{G}$ differ in size by at most one, so none exceeds $\lceil |\mathcal{G}|/M \rceil$. (d) restates Lemma F.3. For (e), $z \mapsto e_{c(z)}$ depends on position alone, and the positional table is indexed by sequence position, so its entries at separator positions may be prescribed independently of its entries at cell positions. □

*Remark.* Condition (a) ensures correctness. Conditions (b), (c), and (d) provide quantitative bounds on width, logit scale, and grid independence. One can verify Appendix G treating $M$ symbolically; substituting the colorings of §F.3 (used in Section 5) yields concrete widths.

---

**PART III — LAYER 1: THE SPATIAL GATHER**

---

# G LAYER 1: SPATIAL GATHER CONSTRUCTIONS

This section proves the gather component of Theorem 7 by constructing the Layer-1 weights. At any position $n$ with virtual time step $\tau(n) \geq 1$, Layer 1 must recover the preceding states

$$\{x_{\tau(n)-1,\ \kappa(n)+\delta} : \delta \in \mathcal{U}\}$$

and track their offsets $\delta$. Localizing these states in separate residual blocks allows Layer 2 to read $D_Q$ at the current token and $D_K$ at an earlier token. Formulating this requirement via $(\tau(n), \kappa(n))$ ensures it covers both cells and separators, and retains cells with unresolved queries as necessary keys (Remark G.5).

Two challenges arise: routing the cells in $\mathcal{U}$ to the query token, and preserving their spatial identities. We provide two constructions resolving both:

(I) **Dedicated-heads.** Each offset $\delta \in \mathcal{U}$ receives its own attention head. The head directly writes the state at offset $\delta$ into a dedicated residual block, so attention performs both routing and separation by itself.

(II) **Shared-head.** A single head attends jointly to all offsets in $\mathcal{U}$. Since the head returns only one weighted sum, the routed states would normally be mixed together. We avoid this by encoding the cell state in the attention *weight* and the routed-cell identity as a distinct basis vector in the value projection using a positional coloring. A one-hidden-layer MLP then separates the aggregate and reconstructs one state block per offset.

Both constructions rely on the spatial-temporal addressing primitive developed above; Construction II also requires the coloring primitive. We first recall the necessary properties.

Throughout, $\mathcal{G}$ denotes the $d$-dimensional torus of volume $|\mathcal{G}|$ (Definition 1), evaluated for $T$ time steps. For a stream position $n$, $\kappa(n)$ and $\tau(n)$ denote its virtual cell and virtual time step (Appendix C), and $\hat{u}_d$ denotes the successor offset. See Table 3 for further notation.

**Spatial-temporal addressing.** For any set of offsets $\mathcal{N}$, Corollary E.6 provides a positional table $p_n \in \mathbb{R}^{d_p}$ and projections $W_Q^{(1)}, W_K^{(1)}$ which, at unit logit scale, score the keys at cells $\kappa(n)+\delta$, $\delta \in \mathcal{N}$, in the previous time step strictly above every other causally visible cell key, with a margin $\gamma > 0$. Separator keys are covered by the gate of §G.2. Its positional score is

$$\langle W_Q^{(1)} p_n,\ W_K^{(1)} p_{n'} \rangle = g\big(\kappa(n) + \hat{u}_d - \kappa(n')\big) + \lambda h\big(\tau(n) - \tau(n')\big),$$

where $g$ favors the requested spatial offsets and $h$ selects temporal displacement 1. The spatial width is $d_S = 2d$ for a single offset and $d_S = O(|\mathcal{N}|\, d \log |\mathcal{N}|)$ in general, and it is independent of the grid volume and the horizon. After multiplying the logits by $\Lambda > 0$, Corollary E.5 bounds the attention mass outside the target set by $N_{\neq} e^{-\Lambda\gamma}$, where $N_{\neq}$ is the number of visible non-target keys.

**Positional colors.** Construction II must additionally preserve the identities of the cells routed by the Shared-head. Corollary F.5 provides a coloring $c$ of $\mathcal{G}$ which is injective on every translate $z + \mathcal{U}$, so the routing cells $\{z + \delta : \delta \in \mathcal{U}\}$ carry pairwise distinct colors at every query location $z$. Writing $M$ for the number of colors, such a coloring exists with

$$M \le \big|(\mathcal{U} - \mathcal{U}) \setminus \{0\}\big| + 1 \le |\mathcal{U}|^2 - |\mathcal{U}| + 1,$$

a bound depending on $\mathcal{U}$ and on nothing else. The same corollary provides the positional storage for the query-side color addresses $e_{c(z+\delta)}$, the key-side own-color block, and a bound on how many causally visible cells share one color. Construction II uses all four in §G.4.

### G.1 Target representation and residual-stream layout

Both Layer-1 constructions produce the same output: $|\mathcal{U}|$ state blocks. For each $\delta \in \mathcal{U}$, the $\delta$-th block should represent the state of cell $(t-1, z+\delta)$.

Because finite softmax attention may leave mass on non-target keys, we first formalize how a block approximates the desired state.

**Definition G.1** ($\bar{\varepsilon}$-approximate one-hot)**.** For $\bar{\varepsilon} \in [0, 1)$ and $s \in \mathcal{S}$, a width-$V$ block $P$ is an $\bar{\varepsilon}$-*approximate one-hot* for $e(s)$ if it is a sub-probability vector, meaning that it is non-negative and has total mass at most 1, and it places mass at least $1 - \bar{\varepsilon}$ on coordinate $s$. Consequently, at most $\bar{\varepsilon}$ mass lies on the remaining $V - 1$ coordinates. We call $\bar{\varepsilon}$ the *per-block error*.

**Definition G.2** (Stream layout)**.** Both constructions use a residual stream

$$x_n = [\, B_0; B_1; B_2; B_3; B_4 \,]$$

with the following orthogonal blocks.

- $B_0 \in \mathbb{R}^3$ is the *gate block*. It stores the separator indicator $s_n$, the auxiliary-token marker $\mu_n \coloneqq \mathbf{1}[n = \star]$, and a constant 1.
- $B_1 \in \mathbb{R}^V$ stores the token's own state one-hot and is used as the value source for state retrieval.
- $B_2 \in \mathbb{R}^{d_p}$ stores the positional features. Besides the $d_S + 2$ spatial-temporal addressing coordinates of Corollary E.6, it carries the *initial-step indicator*

  $$\zeta_n \coloneqq \mathbf{1}\big[\tau(n) = 0\big], \tag{33}$$

  one coordinate, used by the Layer-2 key-validity gate of Definition H.1. Construction II augments these with the query-side color addresses and the own-color block (34).
- $B_3$ is the Layer-1 scratch space. It contains

  $$B_3^{\mathrm{pat}} = \bigoplus_{\delta \in \mathcal{U}} \mathbb{R}^V,$$

  whose $\delta$-th block will hold the state routed from offset $\delta$. Construction I writes these blocks directly and uses no additional scratch. Construction II also uses $B_3^{\mathrm{agg}} \in \mathbb{R}^{M+1}$ for the Shared-head aggregate; its MLP decodes this aggregate into $B_3^{\mathrm{pat}}$ and then clears $B_3^{\mathrm{agg}}$.
- $B_4 \in \mathbb{R}^V$ is reserved for the Layer-2 output and is initialized to zero on ordinary tokens.

For an ordinary cell token at $(t, z)$,

$$B_0 = (0, 0, 1), \qquad B_1 = e(x_{t,z}),$$

while a separator has

$$B_0 = (1, 0, 1), \qquad B_1 = 0.$$

Blocks $B_3$ and $B_4$ are initialized to zero. When the auxiliary token $\star$ of §G.2 is present, it has

$$B_0 = (0, 1, 1), \qquad B_1 = 0, \qquad B_2 = 0, \qquad B_3 = 0, \qquad B_4 = 0,$$

so that $s_\star = \zeta_\star = 0$. All other tokens are *ordinary* and have $\mu_n = 0$.

*Remark* G.3 (The initial-step indicator). While the positional table carries the temporal pair $\psi(\tau) = (\cos 2\pi\tau/T, \sin 2\pi\tau/T)$ (Eq. 29), the indicator $\zeta_n = \mathbf{1}[\tau(n) = 0]$ is not a linear functional of it for $T \geq 3$, necessitating a dedicated coordinate. Like all positional features, it depends solely on sequence position and is readily learned.

This indicator resolves a role asymmetry: an initial-step cell carries $x_{0,z}$ (needed by Layer 1 for queries at $t = 1$) but lacks a parent configuration (excluding it as a Layer-2 key, consistent with $t' \geq 1$ in Eq. 8). Consequently, Layer 1 gates only separators, whereas Layer 2 gates both separators and initial-step cells (§H.5).

**Accounting for $B_2$.** Construction I stores only the addressing coordinates, so

$$d_p = d_S + 3, \qquad d_S = \begin{cases} 2d, & |\mathcal{N}| = 1, \\ O\big(|\mathcal{N}|\, d \log |\mathcal{N}|\big), & \text{in general,} \end{cases}$$

independently of the grid volume and the horizon. Two of the three non-spectral coordinates are the temporal pair of Corollary E.6; the third is the initial-step indicator $\zeta_n$ of (33), which Layer 2 needs and which we account for here so that one positional width serves the whole circuit.

Construction II adds two color blocks. For every offset $\delta \in \mathcal{U}$ the table stores the query-side address $e_{c(z+\delta)}$ of the routed cell, which Corollary F.5*(b)* carries in at most

$$|\mathcal{U}|(M-1) + 1$$

coordinates; these addresses let the Layer-1 MLP identify which coordinate of the Shared-head aggregate belongs to each offset. Separately, the Shared-head needs the key-side color of the key's own cell, because its value projection emits exactly that. By Corollary F.5*(e)* the table carries it as a dedicated width-$M$ block

$$B_2^{\mathrm{col}} \coloneqq (1 - s_n)\, e_{c(\kappa(n))} \in \mathbb{R}^M, \tag{34}$$

the own-color one-hot at cell positions and the zero vector at separator positions. We store this block separately from the $\delta = 0$ address. Separate storage makes the construction independent of whether $0 \in \mathcal{U}$, and it makes a separator's Layer-1 value exactly zero, so no separator color can enter the Shared-head aggregate. This block does not admit the "$M - 1$ coordinates plus a shared constant" compression used for the query addresses, because the reconstructed coordinate would read 1, not 0, at a separator; it therefore costs a full $M$ coordinates.

We can now state our goal as an interface condition on $B_3^{\mathrm{pat}}$.

**Definition G.4** (Spatial induction head, quantitatively)**.** Fix $\bar{\varepsilon} \in [0, 1)$. A two-layer circuit with the stream layout of Definition G.2 is a *spatial induction head at per-block error* $\bar{\varepsilon}$ if:

(i) *The gather.* At *every* position $n$ with $\tau(n) \geq 1$, after Layer 1 the pattern scratch $\Pi_{B_3^{\mathrm{pat}}} x_n$ consists of $|\mathcal{U}|$ width-$V$ blocks indexed by $\delta \in \mathcal{U}$, and the $\delta$-th block is an $\bar{\varepsilon}$-approximate one-hot for

$$e\big(x_{\tau(n)-1,\, \kappa(n)+\delta}\big).$$

At positions with $\tau(n) = 0$ only the weaker property of Lemma G.13 is claimed.

(ii) *The match.* Layer 2 compares the $k$ blocks indexed by $D_Q$ at the current query with the $k$ blocks indexed by $D_K$ at earlier tokens, and averages the states carried by the matching precedents themselves, that is their own $B_1$ entries, into $B_4$.

*Remark* G.5 (The gather covers more positions than the prediction). Condition *(i)* quantifies over all positions with $\tau(n) \geq 1$, a superset of the resolvable queries. Layer 2 reads $D_K$-blocks at every valid earlier key, requiring only that the cell records an observed transition. This key set properly contains the query set via two additions:

- *Row-final cells* ($z_d = W_d - 1, t \geq 1$) pose no cell query (their successor is a separator, Lemma C.3), but supply a valid observation $x_{t,z} \sim f(\cdot \mid P_{t,z})$ and must be retained as keys.
- *Cells with unresolved queries* provide valid precedents $P_{t,z}$ regardless of whether their own query $Q_{t,z}$ is resolvable.

Conversely, initial-step cells are excluded from both sets, an exclusion the circuit explicitly implements (Remark G.3). Resolvability restricts the downstream *prediction* (Appendix H), not the *representation* derived here.

In this section, we prove condition *(i)*. In Appendix H, we use only this interface and construct the common Layer-2 match.

### G.2 Separator gating and finite-softmax error

Before giving the two constructions, we isolate two issues common to both. First, separators must be excluded from Layer-1 attention. Second, when a head is intended to select a single routing cell, finite softmax leaves a small amount of mass elsewhere.

Separators need excluding because their addresses duplicate cells'. By Lemma C.4*(i)* a separator carries the virtual index of the cell *one step before its successor*. A `ROW_SEP` closing row $r$ of time step $t$ has

$$\big(\tau(n), \kappa(n)\big) = \big(t,\ (r+1,\ W_d - 1)\big),$$

the last-column cell of the *following* row, and a `TIME_SEP` closing step $t$ has $(\tau(n), \kappa(n)) = (t+1,\ (0, \ldots, 0, W_d - 1))$, the last-column cell of the first row of the *following* step. In either case the separator shares its virtual spatial-temporal address with an ordinary cell, and both are visible to any sufficiently late query, so a purely positional score cannot tell them apart. Without a gate, a query addressing that cell splits its mass with the duplicate separator.

**Proposition G.6** (The Layer-1 separator gate)**.** *Add $-\Gamma_1 s_{n'}$ to the Layer-1 score, implemented by a rank-one term from the constant query coordinate in $B_0$ to the separator indicator in the key. If*

$$\Gamma_1 \ \geq\ \Lambda\,(R + \gamma), \qquad R \coloneqq \max_{\delta \in \mathcal{G}} g(\delta) - \min_{\delta \in \hat{u}_d - \mathcal{N}} g(\delta) \ \geq\ 0, \tag{35}$$

*with $R$ the quantity already fixed in* (26)*, then every separator key scores at least $\Lambda\gamma$ below every target key. Separators are then absorbed into the same off-target bound $N_{\neq} e^{-\Lambda\gamma}$ of Corollary E.5 as the remaining non-target keys.*

*Proof.* Write the ungated score as $\Lambda G(\delta, s)$ in the notation of Stage 6 of Corollary E.6, and put $m \coloneqq \min_{\delta \in \hat{u}_d - \mathcal{N}} g(\delta)$, so that the target floor of case (a) of that stage is $\Lambda(m + \lambda)$. A separator key attains at most

$$\Lambda G \ \leq\ \Lambda\big(m + R + \lambda\big)$$

before the gate: its spatial term is at most $\max_{\mathcal{G}} g = m + R$ by the definition of $R$, and its temporal term is at most $\lambda$, attained when $h = 1$. After subtracting $\Gamma_1 \geq \Lambda(R + \gamma)$ its score is at most

$$\Lambda(m + R + \lambda) - \Lambda(R + \gamma) \ =\ \Lambda(m + \lambda) - \Lambda\gamma,$$

which is at least $\Lambda\gamma$ below the target floor. Separators therefore satisfy the same isolation inequality as the non-target cell keys, and Corollary E.5 applies with $N_{\neq}$ counting both. □

*Remark.* The gate multiplies a token-embedding coordinate which is already present in $B_0$, so it adds no positional coordinate and does not alter the width bounds of the addressing construction. It is query independent and provides no additional spatial resolution. The bound (35) is loose: it charges every separator the worst spatial score on the whole grid. We do not gate initial-step cells here; see Remark G.3.

*Remark.* The weaker choice $\Gamma_1 = \Lambda$ does not support the stated bound. A separator whose address coincides with a target cell's may also score the target floor $\Lambda(m+\lambda)$ of the proof above, so subtracting $\Lambda$ leaves it only $\Lambda$ below that floor, for relative weight $e^{-\Lambda}$, against the $e^{-\Lambda\gamma}$ claimed downstream. The two agree only at $\gamma = 1$, and $\gamma > 1$ occurs on small grids: on the three-cell torus with $T = 3$ the singleton spatial kernel of Lemma E.3*(ii)* has $\gamma = 1 - \cos(2\pi/3) = 3/2$, and at $\Lambda = 10$ and $N = 12$ the true block error $e^{-10} \approx 4.54 \times 10^{-5}$ exceeds the claimed $Ne^{-\Lambda\gamma} \approx 3.67 \times 10^{-6}$.

**The auxiliary token.** Construction II always, and Construction I when a Dirichlet prior is used, prepend the auxiliary token $\star$ of Definition G.2 at position 0. Every term involving $\star$ is either a rank-one product of the query constant $1 \in B_0$ with the key marker $\mu_{n'}$, or a value term reading $\mu_{n'}$. Since $\mu$ vanishes at ordinary tokens, these terms leave every ordinary score and value unchanged.

Prepending $\star$ shifts every ordinary position by one, so the layout becomes $\rho_\star(t,z) = \rho(t,z)+1$ and the stream length becomes $N_\star = N+1$. Each ordinary token keeps the positional features it had before the shift, now stored at column $\rho_\star$ of the table, so the virtual indices $\kappa$ and $\tau$ and every ordinary score are unchanged. When $\star$ is present, every bound below that involves $N$ is read with $N_\star$ in its place.

**Lemma G.7** (Layer-1 role of $\star$)**.** *At the key $\star$, the positional score vanishes because $B_2 = 0$, the content score vanishes because $B_1 = 0$, and the separator gate vanishes because $s_\star = 0$. Hence:*

*(a)* Construction II. *The term $\Lambda g_\star \cdot 1 \cdot \mu_{n'}$ in $W_Q^{(1)\top} W_K^{(1)}$ gives $\star$ the logit $\Lambda g_\star$ at every query. The value term $e_\star \mu_{n'}$, with $e_\star$ the last coordinate of $B_3^{\mathrm{agg}} \in \mathbb{R}^{M+1}$, gives $\star$ the value $e_\star$; ordinary keys contribute nothing to that coordinate.*

*(b)* Construction I. *Let $\Lambda(m_\delta + \lambda)$ be the target floor of head $\delta$ (Stage 6 of Corollary E.6 at $\mathcal{N} = \{\delta\}$). The term $\Lambda(m_\delta + \lambda - \gamma) \cdot 1 \cdot \mu_{n'}$ in head $\delta$ places $\star$ exactly $\Lambda\gamma$ below the target, and its value $\iota_\delta \Pi_{B_1} x_\star$ is zero. Thus $\star$ is a non-target key covered by Corollary E.5.*

*(c)* As a query, *$\star$ attends only to itself, and its pattern scratch $B_3^{\mathrm{pat}}$ is zero after Layer 1: in Construction I every head returns $\iota_\delta \Pi_{B_1} x_\star = 0$, and in Construction II this is shown in Step 5 of the proof of Theorem G.10.*

**Lemma G.8** (Per-block error of a single head)**.** *Suppose a Layer-1 head places attention mass at least $1-\bar{\varepsilon}$ on the routing cell $(t-1, z+\delta)$ and reads values from $B_1$ into one width-$V$ block of $B_3^{\mathrm{pat}}$. Then that block is an $\bar{\varepsilon}$-approximate one-hot for $e(x_{t-1,z+\delta})$. In particular, Corollary E.5 guarantees*

$$\bar{\varepsilon} \le N e^{-\Lambda\gamma},$$

*and hence any prescribed $\bar{\varepsilon} > 0$ is achieved whenever*

$$\Lambda \ge \gamma^{-1} \log(N/\bar{\varepsilon}).$$

*Proof.* The written block is a non-negative weighted sum of the state one-hots carried by visible cell tokens. The addressed routing cell contributes mass at least $1-\bar{\varepsilon}$ to its correct state coordinate, while all remaining attention mass is at most $\bar{\varepsilon}$, so the block places at most $\bar{\varepsilon}$ on the other $V-1$ coordinates.

Three kinds of key make up that remaining mass, and all are covered. Ordinary cell keys at non-target offsets are $\Lambda\gamma$ below the target by isolation. Separator keys are $\Lambda\gamma$ below it by Proposition G.6, and in addition contribute the zero vector, since a separator has $B_1 = 0$; they can only lower the block's total mass, never move it to a wrong coordinate. The key $\star$, when present, is $\Lambda\gamma$ below the target and contributes the zero vector by Lemma G.7*(b)*. The block is therefore non-negative with total mass at most 1, so Definition G.1 asks only for a sub-probability vector.

For the scale bound, apply Corollary E.5 with the target set the single addressed cell. Every other visible position is a non-target key, so $N_{\neq} \le N-1 \le N$, giving $\bar{\varepsilon} \le N e^{-\Lambda\gamma}$ and hence the stated condition on $\Lambda$. □

### G.3 Construction I: One head per offset

In the first construction, head $\delta$ addresses one offset and writes its state into the block reserved for that same offset, so routing and separation happen together.

**Theorem G.9** (Dedicated-heads gather)**.** *Build Layer 1 with $h = |\mathcal{U}|$ heads, one head for each $\delta \in \mathcal{U}$, and no MLP, incorporating the separator gate of Proposition G.6. At logit scale $\Lambda$, this realizes condition* (i) *of Definition G.4 with*

$$\bar{\varepsilon} \le N e^{-\Lambda\gamma},$$

*and with*

$$d_{\mathrm{MLP}} = 0, \qquad d_{\mathrm{model}} = (|\mathcal{U}|+2)V + 2d + 6, \qquad d_p = 2d+3.$$

*The same holds with the auxiliary token $\star$ configured as in Lemma G.7*(b)*.*

*Proof.* Fix $\delta \in \mathcal{U}$. Head $\delta$ must reach the key at cell $\kappa(n)+\delta$, so it applies Corollary E.6 to the singleton offset set

$$\mathcal{N}_\delta \coloneqq \{\delta\}.$$

For a singleton the spatial addressing width is $d_S = 2d$ by Lemma E.3*(ii)*, independently of which singleton it is, so all $|\mathcal{U}|$ heads share one spatial table of width $2d$ and differ only through their query-key projections, which rotate the $2 \times 2$ blocks (15) by the angle belonging to their own offset. With the two temporal coordinates and the indicator $\zeta_n$ of (33), the positional width is $2d + 3$.

At a position $n$ with $\tau(n) \geq 1$, head $\delta$ therefore places all but at most $Ne^{-\Lambda\gamma}$ of its attention mass on the cell $(\tau(n) - 1,\ \kappa(n) + \delta)$, using Proposition G.6 to cover separator keys. Its value projection is

$$W_V^{(1),\delta} = \iota_\delta \circ \Pi_{B_1},$$

where $\iota_\delta$ embeds $\mathbb{R}^V$ into the $\delta$-indexed block of $B_3^{\mathrm{pat}}$. By Lemma G.8, that block is an $\bar{\varepsilon}$-approximate one-hot for $e(x_{\tau(n)-1,\,\kappa(n)+\delta})$ with $\bar{\varepsilon} \leq Ne^{-\Lambda\gamma}$.

This argument depends on the query solely through $\kappa(n)$ and $\tau(n)$, applying to every position with $\tau(n) \geq 1$ (both cells and separators) as required by Definition G.4(i). The addressed cells are causally visible in all cases: step $t - 1$ strictly precedes any cell token at step $t \geq 1$ (since $\rho$ is order-preserving), and Lemma C.4(ii) establishes the same for separators. This guarantees coverage for row-final cells and cells with unresolved queries, both necessary for Layer 2 (Remark G.5).

Repeating this independently for every $\delta \in \mathcal{U}$ produces all required blocks. Because the heads write to disjoint blocks, no Layer-1 MLP is needed. The residual width is therefore

$$\underbrace{3}_{B_0} + \underbrace{V}_{B_1} + \underbrace{(2d+3)}_{B_2} + \underbrace{|\mathcal{U}|V}_{B_3} + \underbrace{V}_{B_4} = (|\mathcal{U}| + 2)V + 2d + 6.$$

Thus condition *(i)* of Definition G.4 holds. □

### G.4 Construction II: A shared head with MLP decoding

Construction II uses a single Layer-1 head to serve all offsets in $\mathcal{U}$. Since an attention head outputs a single weighted sum, attending to all routed cells simultaneously would normally destroy their spatial identities.

The construction resolves this by splitting the information across *weights* and *values*. The state of a routed cell controls its scalar attention weight, while its positional color determines the direction of its value vector. Because the routing cells have distinct colors, they occupy distinct coordinates of the Shared-head aggregate. A fixed reference token carries a known weight through the same softmax, so the MLP can cancel the shared normalization, and the query stores the color expected at each offset. The MLP can therefore select the coordinate belonging to a given offset and recover its discrete state by threshold comparisons.

The construction critically relies on the discrete state alphabet. Consecutive states differ by 1, so a half-integer threshold is separated by $1/2$ from any valid state. The MLP recovers the exact state provided contamination from non-routing cells remains strictly below this margin, alleviating the need for perfectly sparse attention.

**Layer-1 reference token.** Construction II uses the auxiliary token $\star$ of Lemma G.7*(a)* as a Layer-1 reference for the shared softmax normalization: its logit $\Lambda g_\star$ is set by the marker term and so carries no temporal factor, and its value is the dedicated coordinate $e_\star$. Its Layer-2 role is fixed in Appendix H.

**Theorem G.10** (Shared-head gather)**.** *Let Layer 1 have a single head whose value is the own-color block* (34)*, for a coloring $c$ admissible for $\mathcal{U}$ with $M$ colors in the sense of Definition F.1; in particular an ordinary cell $z'$ emits $e_{c(z')}$ and a separator emits* 0*. Let $\gamma$ be the isolation margin of Corollary E.6 for $\mathcal{N} = \mathcal{U}$, and let $n_{\max}$ be the largest number of causally visible cells carrying one color. By Corollary F.5*(c)*, $n_{\max} \leq T \max_s |c^{-1}(s)|$, and $n_{\max} \leq T\lceil |\mathcal{G}|/M \rceil$ if $c$ is equitable. If*

$$\Lambda\gamma \geq \log\big(2V(n_{\max} - 1)\big), \tag{36}$$

*then a one-hidden-layer ReLU MLP of width*

$$d_{\mathrm{MLP}} \leq 2|\mathcal{U}|M(V-1) + M + 2$$

*recovers $e(x_{\tau(n)-1,\,\kappa(n)+\delta})$ for every $\delta \in \mathcal{U}$, at every position with $\tau(n) \geq 1$. Hence condition* (i) *of Definition G.4 holds exactly, with $\bar{\varepsilon} = 0$.*

*The positional and residual widths are*

$$d_p \le \underbrace{d_S + 3}_{\text{addressing, } \zeta} + \underbrace{|\mathcal{U}|(M-1)+1}_{\text{query addresses}} + \underbrace{M}_{B_2^{\mathrm{col}}}, \qquad d_S = O\big(|\mathcal{U}|\, d \log |\mathcal{U}|\big),$$

*and*

$$d_{\mathrm{model}} = (|\mathcal{U}|+2)V + d_p + M + 4.$$

*Proof.* The computation is implemented in five sequential stages.

Throughout, write $z \coloneqq \kappa(n)$ and $t \coloneqq \tau(n)$ for the query's virtual cell and time step, so that the argument applies verbatim at separator queries as well as at cell tokens.

**1. Encode the state in the attention weight.** At the query $n$, the Layer-1 logit of a causally visible key $n'$ is

$$\ell(n,n') = \underbrace{\Lambda\Big[g\big(z+\hat{u}_d-\kappa(n')\big)+\lambda\, h\big(t-\tau(n')\big)\Big]-\Gamma_1 s_{n'}}_{\text{positional}} + \underbrace{\big\langle w, \Pi_{B_1} x_{n'}\big\rangle}_{\text{content}}, \qquad w_v = \log(v{+}1). \tag{37}$$

The positional part is the spatiotemporal score of Corollary E.6 at $\mathcal{N} = \mathcal{U}$, which is large exactly on the routing cells $\kappa(n') = z + \delta$, $\delta \in \mathcal{U}$, in step $t-1$, together with the separator gate of Proposition G.6. The content part is linear in the one-hot state, equal to $\log(x_{\tau(n'),\kappa(n')}+1)$ at a cell key, which is $\log(x_{t-1,\kappa(n')}+1)$ at a routed cell, and to 0 at a separator, so block-diagonal query-key projections realize (37) without mixing positional and state coordinates.

Write $w_{n'}$ for the exponentiated positional part of (37). Every key in step $t-1$ has $h=1$ and so carries the same temporal factor $e^{\Lambda\lambda}$; for a routing cell at offset $\delta \in \mathcal{U}$ we absorb that factor into

$$w_\delta \coloneqq e^{\Lambda\left(g(\hat{u}_d-\delta)+\lambda\right)}, \tag{38}$$

the positional weight the kernel assigns to the key at offset $\delta$. Let $Z$ denote the softmax partition function over all causally visible keys. The attention weight assigned to the routed cell at offset $\delta$ is then

$$\frac{(x_{t-1,z+\delta}+1)w_\delta}{Z},$$

so the state has been encoded as a scalar factor in the attention weight.

Two bookkeeping points carry through the remaining steps. Each $w_\delta$ is a fixed constant per offset, known when the weights are set, and it carries $e^{\Lambda\lambda}$ inside it; Steps 3 and 4 use $w_\delta$ only through the ratios $w_\delta/w_\star$ against the reference weight and $Z_{\max}/w_\delta$, both fixed at the same time. The partition function $Z$ is the one quantity here which the weights cannot know, and the reference coordinate of Step 3 removes it.

**2. Encode routed-cell identity in the value.** The value projection reads the own-color block (34),

$$W_V^{(1)} = \Pi_{B_2^{\mathrm{col}}},$$

so an ordinary cell $z'$ outputs its color code $e_{c(z')} \in \mathbb{R}^M$ and a separator outputs the zero vector, the block being $(1-s_{n'})e_{c(\kappa(n'))}$. Separator keys therefore contribute nothing to the aggregate, so this construction attains $\bar{\varepsilon} = 0$ exactly, with no separator residue. The reference token outputs an additional basis vector $e_\star$, so the Shared-head aggregate is

$$a \in \mathbb{R}^{M+1}.$$

Because the routed cells $z+\delta$, $\delta \in \mathcal{U}$, have distinct colors, the contribution of each routed cell lands in a different coordinate:

$$a[c(z+\delta)] = \frac{(x_{t-1,z+\delta}+1)w_\delta}{Z} + r_\delta, \qquad a[\star] = \frac{w_\star}{Z},$$

where

$$w_\star \coloneqq e^{\Lambda g_\star}$$

and $r_\delta$ is the total contribution from non-routing cells which happen to share the color $c(z+\delta)$.

Ignoring $r_\delta$ for the moment, the desired state is already present in the scalar $a[c(z+\delta)]$. Two obstacles remain: the common normalization $Z$ is unknown to the MLP, and the MLP must select the color coordinate corresponding to offset $\delta$.

**3. Bound contamination and cancel the unknown normalization.** Admissibility guarantees distinct colors only among the routing cells, so non-routing *cells* may still contribute to the same color coordinate; separators cannot, by Step 2. At most $n_{\max} - 1$ visible non-routing cells share any given color, where $n_{\max}$ is the bound of Corollary F.5*(c)*. Each such cell contributes a content factor at most $V$.

Its positional weight $w_{n'}$ is smaller than every routing cell's by at least $e^{-\Lambda\gamma}$. This is Stage 6 of the proof of Corollary E.6, which puts all three kinds of non-routing key at least $\Lambda\gamma$ below the target floor $\Lambda(m+\lambda)$, where $m \coloneqq \min_{\delta\in\mathcal{U}} g(\hat{u}_d - \delta)$: a key in step $t-1$ at an offset outside $\mathcal{U}$ by the spatial margin $\gamma$, a key in any other time step by the temporal margin, since $\lambda\gamma_T \geq R + \gamma$ by (26), and a separator key by Proposition G.6. Excluding the wrong-time keys is the temporal term's duty, and (38) carries that term. Writing $w_{\min} \coloneqq \min_{\delta\in\mathcal{U}} w_\delta = e^{\Lambda(m+\lambda)}$, every non-routing key has $w_{n'} \leq e^{-\Lambda\gamma} w_{\min}$, and hence $w_{n'} \leq e^{-\Lambda\gamma} w_\delta$ for every $\delta \in \mathcal{U}$. Summing over the sharing cells,

$$Zr_\delta \leq V(n_{\max} - 1)e^{-\Lambda\gamma} w_\delta.$$

Under (36),

$$Zr_\delta \leq \frac{1}{2} w_\delta. \tag{39}$$

The reference coordinate removes the unknown partition function. For an offset $\delta$, color slot $s$, and threshold $v \in \{1, \ldots, V-1\}$, define

$$C_{\delta,s,v} \coloneqq a[s] - \Big(v + \frac{1}{2}\Big)\frac{w_\delta}{w_\star} a[\star].$$

At the target color $s = c(z+\delta)$,

$$C_{\delta,c(z+\delta),v} = \frac{w_\delta}{Z}\Big(x_{t-1,z+\delta} - v + \frac{1}{2}\Big) + r_\delta.$$

Thus, apart from the controlled contamination $r_\delta$, the sign of $C_{\delta,c(z+\delta),v}$ tells us whether $x_{t-1,z+\delta} \geq v$.

**4. Use the query's color address to select the correct slot and threshold it.** The query token stores $g_\delta(z) = e_{c(z+\delta)}$ in its positional block $B_2$; thus $g_\delta(z)[s] = 1$ exactly at the color slot for offset $\delta$.

Selecting the slot and thresholding its scalar are nonlinear operations. Since the architecture (§3.3) possesses only a single hidden ReLU layer, we implement both within the same hidden units. Letting $Z_{\max} \geq Z$ be a strict upper bound on the partition function, we define

$$P_{\delta,s,v} \coloneqq G_\delta C_{\delta,s,v} - B_\delta\big(1 - g_\delta(z)[s]\big), \qquad G_\delta \coloneqq \frac{4Z_{\max}}{w_\delta}, \qquad B_\delta \coloneqq G_\delta. \tag{40}$$

Use the pair of hidden units

$$\mathrm{ReLU}(P_{\delta,s,v}), \qquad \mathrm{ReLU}(P_{\delta,s,v} - 1),$$

whose difference is the clamp

$$\chi_{\delta,s,v} = \min\{1, \max\{0, P_{\delta,s,v}\}\}.$$

We claim that

$$\sum_{s=1}^{M} \chi_{\delta,s,v} = \mathbf{1}[x_{t-1,z+\delta} \geq v].$$

First consider a wrong color slot, for which $g_\delta(z)[s] = 0$. Since $a[s] \leq 1$, we have $C_{\delta,s,v} \leq 1$, so

$$P_{\delta,s,v} \leq G_\delta - B_\delta = 0.$$

Thus every wrong slot contributes zero.

At the target slot, the penalty term vanishes. If $x_{t-1,z+\delta} \geq v$, then

$$x_{t-1,z+\delta} - v + \frac{1}{2} \geq \frac{1}{2},$$

and $r_\delta \geq 0$, so

$$P_{\delta,c(z+\delta),v} \geq \frac{4Z_{\max}}{w_\delta}\frac{w_\delta}{2Z} = 2\frac{Z_{\max}}{Z} \geq 2.$$

The clamp therefore equals 1.

If $x_{t-1,z+\delta} < v$, then

$$x_{t-1,z+\delta} - v + \frac{1}{2} \leq -\frac{1}{2}.$$

Using (39),

$$C_{\delta,c(z+\delta),v} \leq -\frac{w_\delta}{2Z} + r_\delta \leq 0,$$

so the clamp evaluates to 0. Thus, the MLP recovers every threshold indicator exactly.

**5. Recover the one-hot state and clear the temporary aggregate.** Threshold indicators determine the discrete state. With the convention $\chi_{\delta,s,V} \coloneqq 0$, coordinate $v \in \{1,\dots,V-1\}$ of the $\delta$-th pattern block receives

$$\sum_{s=1}^{M}\big(\chi_{\delta,s,v} - \chi_{\delta,s,v+1}\big),$$

while coordinate 0 receives

$$1 - \sum_{s=1}^{M}\chi_{\delta,s,1}.$$

These are exactly the coordinates of $e(x_{t-1,z+\delta}) = e(x_{\tau(n)-1,\,\kappa(n)+\delta})$. The constant is supplied by the hidden unit $\mathrm{ReLU}(1-\mu_n)$, which reads the constant and marker coordinates of $B_0$; it equals 1 at ordinary tokens and 0 at $\star$.

At $\star$ the pattern scratch is zero. Since $B_2 = 0$ there, the shared constant of Corollary F.5*(b)* vanishes and $g_\delta(\star) = 0$. Since $\star$ attends only to itself, $a = e_\star$, so $C_{\delta,s,v} = -(v+\frac{1}{2})w_\delta/w_\star < 0$ and $P_{\delta,s,v} < 0$ for all $\delta, s, v$. Every threshold unit and the constant unit therefore vanish at $\star$.

Because $z$ and $t$ denote the query's virtual cell and time step, recovery holds at every position with $\tau(n) \geq 1$ (both cells and separators). Routed cells reside in step $\tau(n)-1$ and are causally visible due to $\rho$'s order preservation ($t \geq 1$) and Lemma C.4(ii) (separators). This fulfills Definition G.4(i) and Remark G.5 entirely.

The Shared-head aggregate $a$ is temporary. As $a$ is non-negative, $M+1$ additional hidden units pass its coordinates unchanged through ReLU and write $-a$ back to $B_3^{\mathrm{agg}}$. The residual connection clears $B_3^{\mathrm{agg}}$, leaving the exact state one-hots in $B_3^{\mathrm{pat}}$.

Finally, the threshold bank uses

$$2|\mathcal{U}|M(V-1)$$

hidden units, aggregate cancellation uses $M+1$, and the constant uses one, giving

$$d_{\mathrm{MLP}} \leq 2|\mathcal{U}|M(V-1) + M + 2.$$

The positional table consists of the $d_S + 2$ spatial-temporal coordinates, the initial-step indicator $\zeta_n$ of (33), the $|\mathcal{U}|(M-1)+1$ query-side color-address coordinates, and the width-$M$ own-color block (34), so

$$d_p \leq d_S + 3 + |\mathcal{U}|(M-1) + 1 + M.$$

The residual stream contains

$$3 + V + d_p + \big(|\mathcal{U}|V + M + 1\big) + V$$

coordinates, yielding

$$d_{\mathrm{model}} = (|\mathcal{U}|+2)V + d_p + M + 4.$$

Thus all $|\mathcal{U}|$ pattern blocks are recovered exactly and condition *(i)* of Definition G.4 holds with $\bar{\varepsilon} = 0$. □

*Remark.* The factor $M$ in the one-hidden-layer width stems from performing color selection and state thresholding within a single nonlinear stage. A two-hidden-layer architecture could decouple these (selecting the color first, then thresholding), reducing the width to

$$|\mathcal{U}|(M + 2V - 2).$$

We present the wider one-hidden-layer construction to match the architecture in §3.3.

**Corollary G.11** (Grid-independent width bounds)**.** *Take $c$ to be the equitable degree-bound coloring supplied by Corollary F.5*(d)*, so that*

$$M \le |\mathcal{U}|^2 - |\mathcal{U}| + 1 \qquad \textit{and} \qquad n_{\max} \le T\lceil |\mathcal{G}|/M \rceil.$$

*Then Theorem G.10 holds with*

$$d_{\mathrm{MLP}} \le 2|\mathcal{U}|\big(|\mathcal{U}|^2 - |\mathcal{U}| + 1\big)(V - 1) + |\mathcal{U}|^2 - |\mathcal{U}| + 3,$$

*and, writing $m := |\mathcal{U}|$,*

$$d_p \le d_S + m^3 - m + 5,$$

*where*

$$d_S = O\big(|\mathcal{U}|\, d \log |\mathcal{U}|\big).$$

*Thus, both widths depend solely on $|\mathcal{U}|$, $V$, and $d$. For the three experimental grids, the periodic colorings of Table 6 yield smaller values than this degree bound; Theorem G.10 holds with those tighter $M$ values.*

*Proof.* We obtain the MLP bound by substituting $M \le m^2 - m + 1$ into Theorem G.10. For the positional bound, the color coordinates of Theorem G.10 total $m(M-1) + M = (m+1)M - m$, which is increasing in $M$, so at $M = m^2 - m + 1$ the sum-of-cubes identity $(m+1)(m^2 - m + 1) = m^3 + 1$ gives

$$(m+1)M - m \;\le\; m^3 + 1 - m.$$

Adding the $d_S + 3$ addressing and indicator coordinates and the single shared constant already counted in $m(M-1) + 1$ yields $d_p \le d_S + m^3 - m + 5$. □

*Remark.* We state the bound $m^3 - m + 5$ for precision: since $m \ge 2$, $m^3 - m + 5 \le m^3 + 3$ (equality at $m = 2$). Here $m \ge 2$ follows from $W_d \ge 2$ in Definition 1: either $\mathcal{U} \supsetneq D_K$, or $D_K + \hat{u}_d = D_K$, in which case $\mathcal{U} = D_K$ is a union of cosets of $\langle \hat{u}_d \rangle$, each of size $W_d \ge 2$. The own-color and indicator coordinates are therefore absorbed by the grid-independent form in Theorem 7((II)).

### G.5 The spatial gather interface

The two constructions differ only in how they produce the same Layer-1 representation. In Appendix H, we therefore treat them through the following common interface.

**Corollary G.12** (Spatial gather interface)**.** *Either Layer-1 construction realizes condition* (i) *of Definition G.4, at every position $n$ with $\tau(n) \ge 1$ and for every $\delta \in \mathcal{U}$:*

- *Construction I uses $|\mathcal{U}|$ heads and no MLP, with*

  $$\bar{\varepsilon} \le N e^{-\Lambda\gamma},$$

  *at the widths of Theorem G.9.*

- *Construction II uses one head and a one-hidden-layer separating MLP, with*

  $$\bar{\varepsilon} = 0,$$

  *at the widths of Theorem G.10 and the grid-independent bounds of Corollary G.11.*

*Proof.* Immediate from Theorems G.9 and G.10. □

Layer 2 also reads the pattern scratch of invalid keys, including positions with $\tau(n) = 0$, where condition *(i)* makes no claim. The Layer-2 gate requires only the following weaker property there.

Table 8: What the two constructions require on the three experimental settings, at the colorings of Table 6.

| | 1D, $V$=2 | 1D, $V$=3 | 2D von Neumann |
|---|---|---|---|
| *Construction I* (Thm. G.9) | | | |
| heads, $d_{\mathrm{MLP}}$ | 4, 0 | 4, 0 | 8, 0 |
| $d_S = 2d$, $d_p = d_S{+}3$ | 2, 5 | 2, 5 | 4, 7 |
| width of $B_3 = \lvert\mathcal{U}\rvert V$ | 8 | 12 | 16 |
| residual $d_{\mathrm{model}}$ | 20 | 26 | 30 |
| block-error threshold $1/(2k{+}2)$ | $1/8$ | $1/8$ | $1/12$ |
| margin $\gamma$ | 0.076 | 0.019 | 0.5 |
| logit scale $\Lambda \geq$ | 95 | 420 | 18 |
| *Construction II* (Thm. G.10) | | | |
| colors $M$ | 4 | 4 | 12 |
| $d_{\mathrm{MLP}} \leq 2\lvert\mathcal{U}\rvert M(V{-}1) + M + 2$ | 38 | 70 | 206 |
| width of $B_3^{\mathrm{agg}} = M{+}1$ | 5 | 5 | 13 |

**Lemma G.13** (Pattern scratch at every position)**.** *Under either construction, at every ordinary position $n$ other than the final* `TIME_SEP`*, including those with $\tau(n) = 0$, every width-$V$ block of $B_3^{\mathrm{pat}}$ is a sub-probability vector. At $\star$, $B_3^{\mathrm{pat}} = 0$.*

*Proof.* The claim at $\star$ is Lemma G.7*(c)*.

*Construction I.* Block $\delta$ is $\sum_{n'} \pi_{n'} \Pi_{B_1} x_{n'}$ with softmax weights $\pi_{n'} \geq 0$ summing to 1, and each $\Pi_{B_1} x_{n'}$ is a one-hot or zero.

*Construction II.* Fix $\delta$. The wrong-slot bound $P_{\delta,s,v} \leq 0$ of Step 4 uses only $a[s] \leq 1$ and so holds at every position. By Corollary F.5*(b)*, $g_\delta(z)$ is a one-hot at every ordinary position, so at most one slot $s_\delta$ has $\chi_{\delta,s,v} \neq 0$; write $\chi_v \coloneqq \chi_{\delta,s_\delta,v}$ and $\chi_V \coloneqq 0$. Since $a[\star] = w_\star/Z > 0$ and $w_\delta/w_\star > 0$, $C_{\delta,s_\delta,v}$ is nonincreasing in $v$; $P$ is increasing in $C$ and the clamp is nondecreasing, so $1 \geq \chi_1 \geq \chi_2 \geq \cdots \geq \chi_V = 0$. Hence coordinate $v \geq 1$ of the block, $\chi_v - \chi_{v+1}$, is non-negative; coordinate 0, $1 - \chi_1$, is non-negative; and the coordinates telescope to $(1 - \chi_1) + (\chi_1 - \chi_V) = 1$. The block is therefore a probability vector. □

### G.6 Construction sizes for the experimental settings

Table 8 evaluates the two constructions on the three experimental settings of Appendix B.4, at the colorings of Table 6. Three entries need comment.

First, the addressing dimension for Construction II is omitted because the bound of Theorem E.4 is vacuous on these grids: at $\lvert\mathcal{U}\rvert = 8$ and $d = 2$ it exceeds $\lvert\widehat{\mathcal{G}}\rvert = 36$. Four characters suffice to isolate $\mathcal{U}$ on $\mathbb{Z}_6^2$ (§E.5), and the general bound is useful when $\lvert\mathcal{U}\rvert\, d \log \lvert\mathcal{U}\rvert$ is much smaller than the volume. For $d = 1$ the bounding box equals the routing set, so Lemma E.3*(i)* applies directly and yields $d_S \leq 2d\lvert\mathcal{U}\rvert = 8$ without randomization.

Second, the $d_{\mathrm{MLP}}$ row reports the one-hidden-layer width $2\lvert\mathcal{U}\rvert M(V-1) + M + 2$ of Theorem G.10. The two-hidden-layer alternative of the remark after that theorem would give $\lvert\mathcal{U}\rvert(M + 2V - 2)$ instead, which is 24 against 38 for 1D $V$=2, 32 against 70 for 1D $V$=3, and 112 against 206 for 2D von Neumann.

Third, the margin $\gamma$ and the logit scale $\Lambda$ belong to Construction I, whose heads address one offset each. There Lemma E.3*(ii)* gives $\gamma = 1 - \cos(2\pi/W_{\max})$, and Lemma G.8 at the block-error threshold on the line above gives $\Lambda \geq \gamma^{-1} \log\big((2k+2)N\big)$. Construction II addresses all of $\mathcal{U}$ with a single head, so it has a different margin and asks for (36) instead; Example E.8 works that case out on $\mathbb{Z}_6^2$.

## Part IV — Layer 2 and end-to-end guarantees

# H LAYER 2: BAYESIAN MATCHING AND END-TO-END GUARANTEES

Per Appendix G, Layer 1 endows each query token with one width-$V$ pattern block per offset $\delta \in \mathcal{U}$. Layer 2 then compares the query configuration at the current token against the parent configuration at every preceding token, retrieving the states that follow matching precedents.

With exact one-hot pattern blocks, this comparison is immediate: the inner product between two aligned $k$-block configurations equals the number of agreeing offsets. An exact match attains score $k$; every non-match scores at most $k-1$. If the softmax isolates exact matches, equal scores compel it to average their state one-hots, yielding the empirical conditional frequency. An additional token injects the Dirichlet pseudocount, transforming this frequency into the posterior predictive.

Finite Layer-1 error perturbs this ideal in two ways: it narrows the score gap between matches and non-matches, and it disperses the scores among true matches, preventing uniform softmax averaging. The Layer-2 scale $\beta$ amplifies both effects: increasing $\beta$ suppresses non-matches but exacerbates dispersion within the matched set.

This section is structured in six parts. First, we formalize the matching head (§H.1). Second, we prove the exact counting and Dirichlet smoothing identity in the zero-error limit (§H.2). Third, we bound the two finite-error perturbations: the loss of separation (§H.3) and the dispersion among matched scores with its predictive cost (§H.4). Fourth, we configure the key-validity gate based on these bounds (§H.5). Fifth, we balance the perturbations to establish the end-to-end spatial induction theorem (§H.6). Finally, we detail the recipe for setting the finite parameters (§H.7).

## H.1 THE CONFIGURATION-MATCHING HEAD

While the Layer-1 representation stores all offsets in $\mathcal{U}$, the Layer-2 query and key require distinct subsets: the current token exposes the parent configuration of its successor (indexed by $D_Q = D_K + \hat{u}_d$), whereas an earlier token exposes its own parent configuration (indexed by $D_K$).

**Definition H.1** (The Layer-2 head)**.** Define the selection projections on $B_3^{\mathrm{pat}}$

$$\Sigma_Q \coloneqq \big(\text{block } \hat{u}_d + \delta\big)_{\delta \in D_K}, \qquad \Sigma_K \coloneqq \big(\text{block } \delta\big)_{\delta \in D_K}.$$

Thus $\Sigma_Q x_{t,z}$ and $\Sigma_K x_{t',z'}$ are both concatenations of $k$ width-$V$ blocks in the same offset order: the former represents the query configuration of the target $z + \hat{u}_d$, and the latter represents the earlier token's parent configuration.

A position $n'$ is a *valid key* if it is a cell token with $\tau(n') \geq 1$; otherwise, it is an *invalid key*. Invalid keys (separators and initial-step cells) record no transitions of $f$: separators carry no state, while initial-step cells lack a defined parent configuration $P_{0,z'}$ and their states are not drawn from $f$. This matches the restriction $t' \geq 1$ on the counts (8). Write

$$\nu_{n'} \coloneqq s_{n'} + \zeta_{n'} \ \in \ \{0,1,2\}, \tag{41}$$

which is 0 at a valid key and at least 1 at an invalid one; both summands are already present in the stream, $s$ in $B_0$ and $\zeta$ in $B_2$ (Definition G.2).

For query scale $\beta > 0$ and key-validity gate $\Gamma_2 > 0$, define

$$W_Q^{(2)} = \beta\, \Sigma_Q \oplus \Pi_{B_0,\mathbf{1}}, \qquad W_K^{(2)} = \Sigma_K \oplus \big(\sigma_\star \Pi_\mu - \Gamma_2 \Pi_\nu\big), \qquad W_V^{(2)} = \Pi_{B_1} + \tfrac{1}{V}\mathbf{1}\cdot\big(\Pi_{B_0,\mathrm{sep}} + \Pi_{B_0,\mu}\big).$$

For a valid key $(t', z')$, the score is

$$\beta\big\langle \Sigma_Q x_{t,z}, \Sigma_K x_{t',z'} \big\rangle,$$

and the value is its state one-hot $e(x_{t',z'})$. For an invalid key, $\Gamma_2 \nu_{n'} \geq \Gamma_2$ is subtracted from this score (Proposition H.7 bounds $\Gamma_2$). The rank-one term in $W_V^{(2)}$ assigns separators and $\star$ the value $V^{-1}\mathbf{1}$ rather than zero, ensuring the head outputs a probability vector (the readout convention of §H.2). The auxiliary token $\star$ is neither a valid nor an invalid key: its pattern scratch is zero (Lemma G.13) and $\nu_\star = 0$, so its score is $\sigma_\star$ at every query. The output is written to $B_4$, without a Layer-2 MLP.

The projections $\Sigma_Q$ and $\Sigma_K$ do not create new residual coordinates: they only select and reorder blocks already present in $B_3^{\mathrm{pat}}$. Neither does the gate, since $\nu$ is the sum of two coordinates the stream already carries; and neither does the value term, which is rank one.

For a fixed query configuration $c = Q_{t,z}$, let $\mathcal{M}_{t,z}$ denote the set of earlier valid keys, that is cell-token positions $(t', z')$ with $t' \geq 1$, whose parent configuration equals $c$. Its cardinality is $n_c$, matching (8) exactly, and $n_{c,s}$ denotes the number of those precedents whose own state is $s \in \mathcal{S}$.

**The score when Layer 1 is exact.** If all selected blocks are exact one-hots, the inner product

$$\langle \Sigma_Q x_{t,z}, \Sigma_K x_{t',z'} \rangle$$

counts the agreeing offsets in $D_K$. Thus, members of $\mathcal{M}_{t,z}$ score $\beta k$, while non-members score at most $\beta(k-1)$. This integer combinatorial margin persists (albeit degraded) even when Layer-1 blocks are approximate.

The gate strength $\Gamma_2$ is set in §H.5, once the matched-score floor and the separation are available.

### H.2 Exact counting and Dirichlet smoothing

We first analyze the ideal computation, where exact matches receive equal scores and all other positions are excluded. In this limit, the attention head acts as a counter.

The following identity governs the hard-attention limit and requires three hypotheses, which an exact gather alone cannot guarantee. While an exact gather ensures the first (equal one-hot configurations yield equal inner products), a finite softmax retains mass everywhere (making the second a strict limit), and the third necessitates a sink token. Theorem H.9 establishes the finite-parameter counterpart, offering an approximation with explicit error bounds.

**Theorem H.2** (Exact Bayesian matching in the hard-attention limit; Proposition 8)**.** *Suppose the Layer-2 head of Definition H.1 satisfies:*

*(H1) it assigns a common score $a$ to all $n_c$ matching precedents in $\mathcal{M}_{t,z}$;*

*(H2) it assigns score $-\infty$ to every other position, non-matching valid keys and invalid keys alike;*

*(H3) for part* (ii)*, the sink carries the prescribed value and score.*

*Then:*

*(i)* Empirical counting. *With value $e(x_{t',z'})$ at precedent $(t', z')$, the head output is*

$$\frac{1}{n_c} \sum_{(t',z') \in \mathcal{M}_{t,z}} e(x_{t',z'}) = \left( \frac{n_{c,s}}{n_c} \right)_{s \in \mathcal{S}}.$$

*(ii)* Dirichlet smoothing. *Fix $\alpha > 0$. Add one auxiliary token $\star$ with value*

$$\frac{1}{V}\mathbf{1}$$

*and score*

$$a + \log(V\alpha).$$

*Then the head output is exactly*

$$\left( \frac{n_{c,s} + \alpha}{n_c + V\alpha} \right)_{s \in \mathcal{S}} = \widehat{P}_\alpha(\cdot \mid Q_{t,z}),$$

*the posterior predictive distribution of Theorem 3.*

*No Layer-2 MLP is required.*

*Proof.* For part (i), equal scores induce a uniform softmax over the $n_c$ matching precedents; averaging their state one-hots yields the normalized state counts.

For part *(ii)*, every matching precedent has unnormalized weight $e^a$, while $\star$ has weight

$$e^{a+\log(V\alpha)} = V\alpha e^a.$$

The total unnormalized mass is therefore $(n_c + V\alpha)e^a$. In coordinate $s$, the matching precedents contribute $n_{c,s}e^a$, while the auxiliary token contributes $V\alpha e^a \cdot (1/V) = \alpha e^a$. Hence

$$\frac{n_{c,s}e^a + \alpha e^a}{(n_c + V\alpha)e^a} = \frac{n_{c,s} + \alpha}{n_c + V\alpha}.$$

□

**The auxiliary token.** The Dirichlet sink is the token $\star$ of Lemma G.7. By Definition H.1, its Layer-2 score is $\sigma_\star$ at every query and its value is $V^{-1}\mathbf{1}$. For $\alpha > 0$ we set $\sigma_\star = a + \log(V\alpha)$. In the deterministic limit $\alpha \to 0$ the Dirichlet role vanishes, but in Construction II the Layer-1 reference role of Lemma G.7*(a)* does not, so $\star$ is retained; we then set $\sigma_\star \le a - \Delta$, and $\star$ enters Theorem H.9*(ii)* as an extraneous position. Construction I includes $\star$ only when $\alpha > 0$, with the Layer-1 score of Lemma G.7*(b)*. In either case $N$ counts $\star$ when it is present, and the Layer-1 guarantees are unchanged by Lemma G.7.

**Readout convention.** Every value used by Layer 2 is a probability vector, by construction. Cell keys, valid or not, carry state one-hots; the sink $\star$ carries the uniform vector $V^{-1}\mathbf{1}$; and separators carry $V^{-1}\mathbf{1}$ as well, by the rank-one term added to $W_V^{(2)}$ in Definition H.1. Had separators kept the value 0 that their empty $B_1$ would otherwise supply, the retrieved vector would have had total mass $1 - \varepsilon_{\mathrm{sep}} < 1$ at every finite gate strength, since a finite softmax places positive weight on every unmasked position, and it could not have been read as a distribution without an explicit renormalization. As constructed, the vector written to $B_4$ is a convex combination of probability vectors and so lies exactly in the probability simplex. For stochastic claims we take

$$\Pi_{B_4} x_n^{(2)}$$

directly as the predictive distribution; we refer to this as the *simplex readout*.

A standard language-model head $\mathrm{softmax}(W_U x_n^{(2)})$ preserves the predicted state for $W_U = \Pi_{B_4}$ (since softmax is monotonic), but generally distorts the full distribution. Consequently, our exact distributional bounds invoke the simplex readout, while deterministic $\arg\max$ statements hold for either.

Theorem H.2 identifies the ideal computation. With finite weights, two departures from that ideal remain: leakage from non-matching positions, and score dispersion among the matches. We bound them separately.

### H.3 Separation of matches from non-matches

The first question is whether approximate pattern blocks still leave a positive score gap between true matches and non-matches.

**Lemma H.3** (Separation of matching configurations)**.** *Let*

$$P = (P_1, \dots, P_k), \qquad P' = (P'_1, \dots, P'_k) \in \mathbb{R}^{kV},$$

*where each $P_h$ is an $\bar\varepsilon$-approximate one-hot for $e(y_h)$ and each $P'_h$ is an $\bar\varepsilon$-approximate one-hot for $e(y'_h)$ in the sense of Definition G.1. Write*

$$y = (y_1, \dots, y_k), \qquad y' = (y'_1, \dots, y'_k).$$

*Then*

$$\langle P, P' \rangle \ge k - 2k\bar\varepsilon \qquad \text{if } y = y',$$

*while*

$$\langle P, P' \rangle \le (k-1) + 2\bar\varepsilon \qquad \text{if } y \ne y'.$$

*Hence the two score ranges are disjoint whenever*

$$\bar\varepsilon < \frac{1}{2k+2}.$$

*Under the Layer-2 scale $\beta$, every matching position then outscores every non-matching position by at least*

$$\Delta := \beta\big(1 - (2k+2)\bar\varepsilon\big).$$

*Proof.* We use two elementary blockwise bounds.

First, every block inner product is at most 1. Indeed, since $P_h$ has total mass at most 1 and every coordinate of $P'_h$ lies in $[0, 1]$,

$$\langle P_h, P'_h \rangle \le \Big(\max_v P'_h[v]\Big) \sum_v P_h[v] \le 1. \tag{42}$$

Now suppose the two blocks represent the same state, $y_h = y'_h$. Keeping only their common target coordinate gives

$$\langle P_h, P'_h \rangle \ge P_h[y_h] P'_h[y_h] \ge (1 - \bar{\varepsilon})^2 \ge 1 - 2\bar{\varepsilon}.$$

If $y = y'$, summing this bound over all $k$ blocks yields

$$\langle P, P' \rangle \ge k - 2k\bar{\varepsilon}.$$

Next suppose $y_h \neq y'_h$ for at least one block $h$. Since $y_h$ is an off-target coordinate for $P'_h$,

$$P'_h[y_h] \le \bar{\varepsilon}.$$

Therefore

$$\begin{aligned} \langle P_h, P'_h \rangle &= P_h[y_h] P'_h[y_h] + \sum_{v \neq y_h} P_h[v] P'_h[v] \\ &\le \bar{\varepsilon} + \sum_{v \neq y_h} P_h[v] \\ &\le 2\bar{\varepsilon}, \end{aligned}$$

because $P_h$ places at most $\bar{\varepsilon}$ total mass away from its own target coordinate. The remaining $k - 1$ blocks contribute at most 1 each by (42), so

$$\langle P, P' \rangle \le (k - 1) + 2\bar{\varepsilon}.$$

The ranges are disjoint exactly when

$$k - 2k\bar{\varepsilon} > (k - 1) + 2\bar{\varepsilon},$$

equivalently $\bar{\varepsilon} < 1/(2k + 2)$. Multiplying the difference by $\beta$ gives the stated gap $\Delta$. □

For deterministic rules, this gap is all we need: every matching precedent carries the same correct next-state value.

**Corollary H.4** (Deterministic matching)**.** *Let the rule be deterministic, let there be $n_c \ge 1$ matching precedents among at most $N$ visible positions, and let $\Delta$ be the gap of Lemma H.3. Assume every other visible position scores at least $\Delta$ below every matching precedent, whether it is a non-matching valid key, a separator, an initial-step cell, or $\star$. Lemma H.3 supplies this for non-matching valid keys; Proposition H.7 of §H.5 supplies it for the invalid keys, and the choice $\sigma_\star \le a - \Delta$ for $\star$; until then the assumption is carried as a hypothesis. Then the Layer-2 head places mass at least $1 - \varepsilon_{\mathrm{ret}}$ on the matching precedents, where*

$$\varepsilon_{\mathrm{ret}} := \frac{(N - n_c)e^{-\Delta}}{n_c + (N - n_c)e^{-\Delta}}.$$

*If $u^\star = e(f(Q_{t,z}))$ is their common value, the output $\hat{y}$ obeys*

$$\|\hat{y} - u^\star\|_2 \le 2\varepsilon_{\mathrm{ret}}.$$

*In particular, if*

$$\Delta > \log\big(2\sqrt{2}\,(N - n_c)/n_c\big),$$

*then the predicted state is exactly correct.*

*Proof.* Let $r^+_{\min}$ be the smallest matched score. By hypothesis every other visible position scores at most $r^+_{\min} - \Delta$, so the total mass $w$ on the matching set satisfies

$$w \ge \frac{n_c e^{r^+_{\min}}}{n_c e^{r^+_{\min}} + (N - n_c)e^{r^+_{\min} - \Delta}} = 1 - \varepsilon_{\mathrm{ret}}.$$

All matching precedents carry the same value $u^\star$, so the head output can be written

$$\hat{y} = wu^\star + (1-w)\rho$$

for some probability vector $\rho$. Hence

$$\|\hat{y} - u^\star\|_2 = (1-w)\|\rho - u^\star\|_2 \le 2(1-w) \le 2\varepsilon_{\mathrm{ret}}.$$

The stated condition makes this perturbation smaller than the distance from $u^\star$ to any decision boundary between two basis vectors, and therefore preserves its $\arg\max$. □

*Remark.* The displayed threshold is only a sufficient condition and is not tight. Its role is to provide a finite explicit scale compatible with the statement of Theorem H.9*(ii)*.

### H.4 Matched-score dispersion and predictive error

For stochastic rules, mere separation is insufficient. Since matching precedents may carry distinct next states, recovering the true counts requires the softmax to remain nearly uniform *within* the matched set. Approximate Layer-1 blocks perturb these matched scores.

**Lemma H.5** (Score dispersion among matches)**.** *In the setting of Lemma H.3, define*

$$a := \beta(k - 2k\bar{\varepsilon}).$$

*Every matched score can be written*

$$a + \eta_j, \qquad 0 \le \eta_j \le \eta, \qquad \eta := 2\beta k\bar{\varepsilon},$$

*and every non-matching score is at most*

$$a - \Delta, \qquad \Delta = \beta\big(1 - (2k+2)\bar{\varepsilon}\big).$$

*Proof.* For a matching pair, Lemma H.3 gives the lower bound $k - 2k\bar{\varepsilon}$, while (42) gives the upper bound $k$. Hence the matched Layer-2 scores lie in

$$\big[\beta(k - 2k\bar{\varepsilon}), \beta k\big],$$

an interval of width $2\beta k\bar{\varepsilon}$. Taking the lower endpoint as $a$ gives the claimed representation $a + \eta_j$.

For a non-match, Lemma H.3 gives score at most

$$\beta\big((k-1) + 2\bar{\varepsilon}\big) = \beta(k - 2k\bar{\varepsilon}) - \beta\big(1 - (2k+2)\bar{\varepsilon}\big) = a - \Delta.$$

□

Layer-1 error thus introduces a tradeoff between:

$$\Delta = \beta\big(1 - (2k+2)\bar{\varepsilon}\big), \qquad \eta = 2\beta k\bar{\varepsilon}.$$

A large $\Delta$ suppresses non-matches, while a small $\eta$ maintains uniform weighting among matches. Increasing $\beta$ scales both simultaneously.

We next quantify the cost of the second effect.

**Lemma H.6** (KL error from matched-score dispersion)**.** *In the setting of Theorem H.2*(ii)*, suppose the matching scores are $a + \eta_j$ with $|\eta_j| \le \eta$, while the sink remains at $a + \log(V\alpha)$. Let $p_\eta$ denote the output after restricting attention to the matching precedents and sink. Then, for every $s \in \mathcal{S}$,*

$$p_\eta(s) \ge e^{-2\eta}\widehat{P}_\alpha(s \mid Q_{t,z}),$$

*and therefore*

$$D_{\mathrm{KL}}\Big(\widehat{P}_\alpha(\cdot \mid Q_{t,z}) \,\Big\|\, p_\eta\Big) \le 2\eta.$$

*No small-$\eta$ assumption is required.*

*Proof.* The $s$-th coordinate of the dispersed matched-plus-sink output is

$$p_\eta(s) = \frac{\sum_{j:x_j=s} e^{\eta_j} + \alpha}{\sum_j e^{\eta_j} + V\alpha}.$$

Because $e^{-\eta} \le e^{\eta_j} \le e^{\eta}$ and $e^{-\eta} \le 1 \le e^{\eta}$, the numerator lies in

$$\big[e^{-\eta}(n_{c,s} + \alpha),\ e^{\eta}(n_{c,s} + \alpha)\big],$$

while the denominator lies in

$$\big[e^{-\eta}(n_c + V\alpha),\ e^{\eta}(n_c + V\alpha)\big].$$

Thus

$$p_\eta(s) \ge e^{-2\eta} \frac{n_{c,s} + \alpha}{n_c + V\alpha} = e^{-2\eta} \widehat{P}_\alpha(s \mid Q_{t,z}).$$

Taking logarithms and averaging under $\widehat{P}_\alpha$ gives the KL bound. □

*Remark.* Lemma H.5 actually gives $\eta_j \ge 0$, in which case the constant can be improved. We keep the symmetric $2\eta$ form because it makes the perturbation lemma independent of how the matched-score interval is anchored and matches the statement used below.

### H.5 The key-validity gate

With the matched-score floor $a$ and the separation $\Delta$ now fixed, we can set the gate strength $\Gamma_2$ which puts every invalid key under the same ceiling as a non-matching valid key.

**Proposition H.7** (The Layer-2 key-validity gate)**.** *In the notation of Lemma H.5, let $a := \beta(k - 2k\bar{\varepsilon})$ be the matched-score floor and $\Delta$ the separation. If*

$$\Gamma_2 \ \ge\ \beta k - a + \Delta \ =\ 2\beta k\bar{\varepsilon} + \Delta, \tag{43}$$

*then every invalid key scores at most $a - \Delta$, the same ceiling that Lemma H.5 gives for non-matching valid keys. Moreover every key's Layer-2 value is a probability vector.*

*Proof. The gate.* An invalid key may carry nonzero pattern scratch, meaning its ungated score is not necessarily small (separators inherit virtual cells and receive Layer-1 gathers, while initial-step cells assemble gathers from the cells they can see). However, (42) bounds the score: every block of $B_3^{\mathrm{pat}}$ is a sub-probability vector at every ordinary position, including those with $\tau(n') = 0$ (Lemma G.13), so each of the $k$ inner products is at most 1, yielding

$$\beta\big\langle \Sigma_Q x_{t,z}, \Sigma_K x_{n'} \big\rangle \ \le\ \beta k =: S_{\max}$$

for every key, valid or not. Since $\nu_{n'} \ge 1$ at an invalid key, its gated score is at most $S_{\max} - \Gamma_2 \le a - \Delta$ by (43).

*The values.* A valid key has $s_{n'} = 0$, so its value is $\Pi_{B_1} x_{n'} = e(x_{t',z'})$, a one-hot. A separator has $s_{n'} = 1$ and $\Pi_{B_1} x_{n'} = 0$, so its value is $V^{-1}\mathbf{1}$. An initial-step cell has $s_{n'} = 0$ and carries its own state one-hot $e(x_{0,z'})$. The token $\star$ has $\mu_\star = 1$ and $\Pi_{B_1} x_\star = 0$, so its value is $V^{-1}\mathbf{1}$. All four are probability vectors. □

*Remark.* Two consequences simplify the remainder of this appendix.

First, invalid keys require no separate error budget. Gated below $a - \Delta$, they become indistinguishable from non-matching valid keys and are absorbed into the $N$ visible positions. Theorem H.9 bounds their mass collectively, eliminating separate $\varepsilon_{\mathrm{sep}}$ bookkeeping.

Second, because all values are probability vectors, the head output (a convex combination) lies exactly in the simplex. The readout convention of §H.2 dictates why separators cannot retain the value 0.

Like the Layer-1 gate, this mechanism uses a rank-one term over existing stream coordinates, avoiding any width increase.

*Remark* H.8 (The initial-step half of the gate is not optional). Gating separators alone falsifies the end-to-end theorem. Take $D_K = \{-1, 0, 1\}, \mathcal{U} = \{-1, 0, 1, 2\}$, $V = 2$, eight cells, and the constant rule $f \equiv 0$, on the trajectory $x_0 = 11111111$, $x_1 = x_2 = 00000000$. Query the first cell of $x_1$: its query configuration is 111, the transition $111 \mapsto 0$ has been observed, so the query is resolvable with $n_c = 1$ and the answer is 0. Every causally visible cell of the initial time step has state 1, so every Layer-1 head returns $e(1)$ there, and each of the eight initial-step tokens presents the spurious parent configuration 111 while carrying its own state 1. Layer 2 sees eight spurious matches valued 1 against one genuine match valued 0, all at the same score, and returns $p(1) = 8/9$. This survives at finite parameters meeting every other condition: at $\Lambda = 100$ and $\beta = 20$ one has $\bar{\varepsilon} \le 5.15 \times 10^{-12}$ and $\Delta \approx 20$, far above the threshold $\log(2\sqrt{2}(N - n_c)/n_c) \approx 4.30$ of Theorem H.9*(ii)*. The $\zeta$ half of $\nu$ in (41) removes the eight spurious keys.

### H.6 End-to-end spatial induction

We now have all the pieces of the finite-parameter argument; Layer 1 fixes the per-block error $\bar{\varepsilon}$, which in turn sets the separation gap $\Delta$ between matches and non-matches and the dispersion $\eta$ within the matched set. The sink token supplies the Dirichlet smoothing. It remains to choose one scale $\beta$ which suppresses the non-matching mass and bounds the matched-set dispersion at the same time.

**Theorem H.9** (End-to-end spatial induction; Theorem 9 of the body)**.** *Let Layer 1 be either construction of Corollary G.12, at per-block error $\bar{\varepsilon}$, with the separator gate of Proposition G.6 at strength $\Gamma_1 \ge \Lambda(R + \gamma)$. Let Layer 2 be the head of Definition H.1 with scale $\beta$ and key-validity gate $\Gamma_2 \ge 2\beta k\bar{\varepsilon} + \Delta$, as in Proposition H.7. Write*

$$\Delta := \beta\big(1 - (2k + 2)\bar{\varepsilon}\big), \qquad a := \beta(k - 2k\bar{\varepsilon}),$$

*and, whenever $\star$ is present, let it be configured as in Lemma G.7 with Layer-2 score $\sigma_\star$.*

*(i)* Stochastic rules. *Fix $\kappa \in (0, 1]$ and $\alpha > 0$, assume $N \ge V\alpha$, and include $\star$ with $\sigma_\star = a + \log(V\alpha)$. If*

$$\Delta \ge \log \frac{4N}{(n_c + V\alpha)\kappa} \qquad \text{and} \qquad \bar{\varepsilon} \le \frac{\kappa}{16k\Delta},$$

*then at every resolvable query, cell token or separator, under the simplex readout, the Layer-2 output $p$ satisfies*

$$D_{\mathrm{KL}}\Big(\widehat{P}_\alpha(\cdot \mid Q_n) \,\Big\|\, p\Big) \le \kappa.$$

*(ii)* Deterministic rules. *Let the rule be deterministic and take $\alpha \to 0$, with $\sigma_\star \le a - \Delta$ whenever $\star$ is present (always in Construction II). If*

$$\bar{\varepsilon} < \frac{1}{2k + 2}$$

*and*

$$\Delta > \log\big(2\sqrt{2}\,(N - n_c)/n_c\big),$$

*then at every resolvable query, cell token or separator, the predicted state is exactly correct. This conclusion holds for either the simplex readout or the standard softmax readout.*

*Remark.* Both gate hypotheses are indispensable. Dropping $\Gamma_1$ to the weaker choice $\Gamma_1 = \Lambda$ breaks the per-block error bound whenever $\gamma > 1$; dropping the $\zeta$ half of $\Gamma_2$ makes part (ii) outright false, by Remark H.8.

*Proof.* Throughout, call a visible position *extraneous* if it is neither a matching precedent nor the sink. Three kinds of position are extraneous: non-matching valid keys, separators, and initial-step cells; in part (ii), where $\star$ plays no Dirichlet role, $\star$ is extraneous as well. We treat them together, and the gate hypotheses enter only here.

*Part (i).* We separate the error into two terms: bias within the matched-plus-sink set and contamination from the extraneous positions.

By Lemma H.5, the matched scores have the form

$$a + \eta_j, \qquad a = \beta(k - 2k\bar{\varepsilon}), \qquad 0 \le \eta_j \le \eta \le 2\beta k\bar{\varepsilon},$$

while every non-matching valid key scores at most $a - \Delta$. By Proposition H.7, the choice of $\Gamma_2$ puts every invalid key under the same ceiling $a - \Delta$. Hence *every* extraneous position scores at most $a - \Delta$, and every one of them carries a probability vector as its value. Place the sink at

$$a + \log(V\alpha).$$

First remove the extraneous positions and renormalize. Let $p_\eta$ denote the resulting matched-plus-sink output. By Lemma H.6,

$$p_\eta \geq e^{-2\eta} \widehat{P}_\alpha$$

coordinatewise.

Now restore them. Let $w$ be their total attention mass and let $R$ be the normalized average of their value vectors, which is a probability vector because each of those values is one. Then

$$p = (1 - w)p_\eta + wR \geq (1 - w)e^{-2\eta} \widehat{P}_\alpha.$$

Therefore

$$D_{\mathrm{KL}}\Big(\widehat{P}_\alpha \,\Big\|\, p\Big) \leq 2\eta - \log(1 - w). \tag{44}$$

We bound the two terms separately.

With at most $N$ visible positions in total, the unnormalized mass of extraneous positions cannot exceed $Ne^{a-\Delta}$. Gating invalid keys proves crucial here: separators and initial-step cells (already counted in $N$) are forced below this $e^{a-\Delta}$ ceiling. Meanwhile, the $n_c$ matching keys each carry weight at least $e^a$, and the sink carries $V\alpha e^a$.

$$w \leq \frac{Ne^{-\Delta}}{n_c + V\alpha} \leq \frac{\kappa}{4}$$

by the first hypothesis. Since $\kappa \leq 1$, we have $w \leq 1/4$ and therefore

$$-\log(1 - w) \leq 2w \leq \frac{\kappa}{2}.$$

For the dispersion term,

$$\beta = \frac{\Delta}{1 - (2k + 2)\bar{\varepsilon}}.$$

The hypotheses imply $\bar{\varepsilon} \leq 1/(4k + 4)$, so the denominator is at least $1/2$ and therefore $\beta \leq 2\Delta$. Consequently

$$\eta \leq 2\beta k\bar{\varepsilon} \leq 4k\Delta\bar{\varepsilon} \leq \frac{\kappa}{4},$$

using the second hypothesis. Thus $2\eta \leq \kappa/2$. Substituting the two bounds into (44) gives the claim.

For completeness, the implication $\bar{\varepsilon} \leq 1/(4k + 4)$ follows from the stated assumptions. Since $n_c \leq N$, $V\alpha \leq N$, and $\kappa \leq 1$, the first hypothesis gives $\Delta \geq \log 2$. Hence

$$\bar{\varepsilon} \leq \frac{1}{16k\Delta} \leq \frac{1}{16k \log 2} \leq \frac{1}{4k + 4}.$$

*Part (ii).* Resolvability gives $n_c \geq 1$, and by the previous paragraphs every extraneous position, invalid keys included, scores at most $a - \Delta$; so does $\star$, since $\sigma_\star \leq a - \Delta$, and its value $V^{-1}\mathbf{1}$ is a probability vector. So Corollary H.4 applies directly with the gap $\Delta$, with its $N - n_c$ non-matching positions read as the extraneous ones, and returns the correct $\arg\max$. For Construction I, choosing the Layer-1 logit scale so that

$$\Lambda\gamma > \log\big((2k + 2)N\big)$$

ensures $\bar{\varepsilon} < 1/(2k + 2)$; Construction II has $\bar{\varepsilon} = 0$ exactly. □

**Separator queries.** The theorem covers separator queries under the same hypotheses. Two conditions are needed. On the query side, Definition G.4*(i)* requires the gather at every position with $\tau(n) \geq 1$, separators included, and Corollary G.12 discharges it for both constructions; the underlying facts are Lemma C.4*(i)–(iii)*, which place the separator's target cell at $\kappa(n) + \hat{u}_d$, its $k$ parents in step $\tau(n) - 1$, and those parents at offsets $D_Q \subseteq \mathcal{U}$ from $\kappa(n)$. On the query side, Layer 2 reads only the $D_Q$ blocks and never the query token's own state, so a separator query needs nothing further, while a separator *key* is gated by $\nu$ (41) and enters the theorem as an extraneous position. A separator is therefore a legitimate query and an illegitimate key, and the single gate $\nu$ separates the two roles. This is the separator entry of Table 5: the same successor rule, positional table and attention heads predict the first cell of a row, so no separate routing construction is required.

### H.7 Choosing finite parameters

The hypotheses of Theorem H.9 couple $\bar{\varepsilon}$, $\Delta$ and $\beta$, and one order of choice satisfies them all. Fix the divergence budget $\kappa \in (0, 1]$ and the pseudocount $\alpha > 0$.

(1) Take $\Delta = \log\big(4N/((n_c + V\alpha)\kappa)\big)$, the smallest value the first hypothesis allows. It depends on the stream length $N$, the smoothing $V\alpha$, the budget $\kappa$ and the match count $n_c$; the paragraph on one circuit below removes the dependence on $n_c$.

(2) Take $\bar{\varepsilon} = \kappa/(16k\Delta)$, the largest per-block error the second hypothesis allows at that $\Delta$.

(3) Take $\beta = \Delta/(1 - (2k+2)\bar{\varepsilon})$, which realizes the prescribed $\Delta$ and, by the proof above, satisfies $\beta \leq 2\Delta$, so the dispersion obeys $\eta \leq \kappa/4$.

(4) Take $\Gamma_1 \geq \Lambda(R + \gamma)$ by Proposition G.6 and $\Gamma_2 \geq 2\beta k\bar{\varepsilon} + \Delta$ by Proposition H.7, and $\sigma_\star = a + \log(V\alpha)$.

The Layer-1 logit scale comes last, and it is where the two constructions differ. Construction II attains $\bar{\varepsilon} = 0$ exactly, so $\eta = 0$, step (2) is vacuous, the dispersion term drops out of (44), and only the leakage term $-\log(1 - w)$ survives; any $\Lambda$ meeting (36) serves. Construction I attains $\bar{\varepsilon} \leq Ne^{-\Lambda\gamma}$, so step (2) becomes a condition on $\Lambda$, namely $\Lambda \geq \gamma^{-1}\log\big(16kN\Delta/\kappa\big)$ by Lemma G.8. For deterministic rules, take $\alpha \to 0$ and replace step (1) by $\Delta > \log\big(2\sqrt{2}(N - n_c)/n_c\big)$, step (2) by $\bar{\varepsilon} < 1/(2k+2)$, and $\sigma_\star$ by any value at most $a - \Delta$; Construction I then needs $\Lambda\gamma > \log\big((2k+2)N\big)$, as in part *(ii)* of the proof.

**One query, or one circuit.** The recipe above reads $n_c$ as the count at the query being bounded, and step (1) gives the smallest $\Delta$ meeting the first hypothesis at that query. The weights of a circuit cannot depend on $n_c$, which is a property of the context and is observed at run time. For one fixed circuit which works uniformly over every resolvable query, run the same recipe at the worst case resolvability permits, $n_c \geq 1$. Step (1) becomes

$$\Delta = \log \frac{4N}{(1 + V\alpha)\kappa},$$

and the deterministic replacement becomes $\Delta > \log\big(2\sqrt{2}\,(N - 1)\big)$; steps (2)–(4) and the Layer-1 logit scale then follow from that $\Delta$ unchanged. The count enters only the first hypothesis, and it enters decreasingly, so meeting that hypothesis at $n_c = 1$ meets it at every resolvable query. No parameter of the model changes with the observed count.

*Remark.* The stochastic inequality characterizes a tradeoff between routing and retrieval. Increasing $\beta$ widens both the separation gap $\Delta$ and the matched-score dispersion $\eta$. Consequently, we cannot lower the softmax temperature arbitrarily. For stochastic rules, an excessive logit scale concentrates mass on a strict subset of matching precedents, destroying the uniform weighting required for empirical counting.

## Part V — Experiments

# I EXPERIMENTAL METHODS AND ADDITIONAL RESULTS

## I.1 DATA GENERATION AND RULE SPLITS

**Trajectory generation.** Each trajectory fixes one rule table across all $T$ time steps. Initial cells ($t$=0) are drawn i.i.d. from $\{0, \ldots, V-1\}$, and all spatial axes use periodic boundary conditions. The neighborhood configuration (left–center–right for 1D; north–west–center–east–south for 2D von Neumann) is base-$V$ encoded to index the rule table. Deterministic rules return the next cell value directly, whereas stochastic rules return a categorical distribution for sampling.

**Serialization and masks.** In 1D, each time step contributes $|\mathcal{G}|$ cell tokens followed by a `TIME_SEP` token (omitting the final separator). In 2D, the grid is serialized row by row, with `ROW_SEP` tokens separating rows and `TIME_SEP` tokens separating time steps. This stream length is one token shorter than $N$ (Definition C.1), but this final omission affects no query cell. Neither the rule table nor rule ID is provided in the input. Training loss spans the full next-token sequence. Validation and test masks isolate the prediction region ($t = T_0, \ldots, T-1$), and reported metrics evaluate only cell-value targets, excluding separators.

**Coverage filtering.** Select evaluation splits apply a trajectory-level coverage filter: candidate trajectories are rejection-sampled until all $V^k$ configurations appear within the $T_0-1$ observed transitions of the context ($t = 0, \ldots, T_0-2$). This inclusion criterion guarantees that every scored configuration has been observed in context. Table 9 details filter usage across splits.

Table 9: Trajectory coverage filter by split. "Yes" means the split rejection-samples trajectories until all $V^k$ neighborhood configurations appear in the context; "No" means no filter. Deterministic runs use five splits (train, val-in without filter, val-in with filter, val-out, test); stochastic runs use three (train, validation, test). The main-text deterministic validation refers to val-out; val-in is diagnostic only and not used for checkpoint selection.

| Dataset type | Train | Val-in (no filter) | Val-in (filter) | Val-out / Val | Test |
|---|---|---|---|---|---|
| Deterministic | No | No | Yes | Yes | Yes |
| Stochastic | No | — | — | Yes | Yes |

**Deterministic rule pools.** Each deterministic trajectory draws one rule uniformly at random (with replacement across trajectories) from the corresponding split's rule pool. Training, validation-out, and test rule pools are disjoint at the canonical-rule level. The split is drawn on canonical rules first, then each canonical rule is expanded through its symmetry orbit, so equivalent rules never appear in different splits.

ECA $V$=2, $k$=3. The 256 elementary rules form 88 canonical classes under reflection, complementation, and their composition. Rules are stratified by Wolfram behavioral classes I–IV before splitting. The canonical-level split is $70/9/9$ (train / val-out / test); expanding each class through its symmetry orbit gives $203/28/25$ element rules. The val-in split draws rules from the training pool with new initial conditions; the val-in-filter variant restricts to training rules that can generate a full-coverage context.

1D $V$=3, $k$=3. The rule space has $3^{27} \approx 7.6 \times 10^{12}$ entries and cannot be enumerated. The equivalence relation uses $S_3$ state permutations; a canonical rule is the lexicographically smallest of its six state-permutation variants. Rules are stratified by sorted output histogram, and the entropy of each histogram defines an entropy bin. Rules are sampled to make the resulting canonical set approximately uniform over discrete entropy values. Pool sizes: 3,621,300 training / 122,640 val-out / 122,640 test rules. For low-entropy strata that rarely admit full-coverage contexts, the val-out and test rules are drawn symmetrically from the coverage-feasible subset; the three canonical pools remain disjoint.

2D VON NEUMANN $V$=2, $k$=5. The rule space has $2^{32} \approx 4.3 \times 10^9$ entries. The equivalence relation uses state complementation only. Rules are stratified by folded Langton $\lambda$-strata (indexed by

$\min(\#1, 32 - \#1)$). Small strata are enumerated; large strata are sampled at fixed number-of-ones per rule. The canonical pools for training, val-out, and test are disjoint. Near-constant strata contribute val-out and test rules symmetrically from the coverage-feasible subset. After orbit expansion and deduplication, the pools contain 315,859 training / 10,904 val-out / 10,904 test rules.

**Stochastic rule sampling.** Each stochastic trajectory uses a fresh rule table fixed across all $T$ steps. For each configuration $c \in [V]^k$, the next-cell categorical distribution $p_c$ is drawn independently from Dirichlet$(1, \ldots, 1)$ over $[V]$. As the Dirichlet distribution is continuous, independent tables coincide with probability zero. Consequently, trajectories almost surely use distinct rules, ensuring validation and test trajectories present rules unobserved during training. Rule-table shapes are $8 \times 2$ (1D $V$=2, $k$=3), $27 \times 3$ (1D $V$=3, $k$=3), and $32 \times 2$ (2D VN $V$=2, $k$=5).

**Dataset sizes.** All settings use 50M training trajectories, 500k validation, and 500k test trajectories. The same counts apply to both deterministic and stochastic datasets. Table 10 summarizes the grid, horizon, and context length for each setting.

Table 10: Grid and trajectory parameters for each experimental setting. Values shown as "det / stoch" when they differ between the two regimes.

| Task | Rule type | Grid $\mathcal{G}$ | $T$ | $T_0$ | $V^k$ configs |
|---|---|---|---|---|---|
| 1D $V$=2, $k$=3 | det / stoch | 16 | 10 / 20 | 4 / 10 | 8 |
| 1D $V$=3, $k$=3 | det / stoch | 32 / 16 | 12 / 20 | 8 / 10 | 27 |
| 2D VN $V$=2, $k$=5 | det / stoch | $6 \times 6$ | 12 / 20 | 8 / 10 | 32 |

### I.2 Models and training

Tables 11 and 12 give the architecture and training hyperparameters for each setting. All models use a two-layer transformer with learned absolute positional embeddings (`nn.Embedding`). The main text refers to the two first-layer designs as **Shared-head** (heads= $[1, 1]$, mlp= $[1, 1]$) and **Dedicated-heads** (heads= $[H, 1]$, mlp= $[0, 1]$).

Table 11: Architecture and training hyperparameters: deterministic settings.

| Task | Arch. | heads | mlp | $d_{\text{model}}$ | $d_{\text{ffn}}$ | lr | wd | $\beta_2$ | ep |
|---|---|---|---|---|---|---|---|---|---|
| 1D $V$=2, $k$=3 | Shared-head | $[1, 1]$ | $[1, 1]$ | 128 | 512 | $10^{-3}$ | 0.2 | 0.999 | 3 |
| | Dedicated-heads | $[4, 1]$ | $[0, 1]$ | 128 | 512 | $10^{-3}$ | 0.2 | 0.999 | 3 |
| 1D $V$=3, $k$=3 | Shared-head | $[1, 1]$ | $[1, 1]$ | 256 | 1024 | $5\times10^{-4}$ | 0.2 | 0.95 | 8 |
| | Dedicated-heads | $[4, 1]$ | $[0, 1]$ | 256 | 1024 | $5\times10^{-4}$ | 0.2 | 0.95 | 8 |
| 2D VN $V$=2, $k$=5 | Shared-head | $[1, 1]$ | $[1, 1]$ | 1024 | 4096 | $10^{-3}$ | 0.2 | 0.95 | 5 |
| | Dedicated-heads | $[8, 1]$ | $[0, 1]$ | 1024 | 4096 | $10^{-3}$ | 0.2 | 0.95 | 5 |

Table 12: Architecture and training hyperparameters: stochastic settings.

| Task | Arch. | heads | mlp | $d_{\text{model}}$ | $d_{\text{ffn}}$ | lr | wd | $\beta_2$ | ep |
|---|---|---|---|---|---|---|---|---|---|
| 1D $V$=2, $k$=3 | Shared-head | $[1, 1]$ | $[1, 1]$ | 256 | 1024 | $10^{-3}$ | 0.2 | 0.999 | 15 |
| | Dedicated-heads | $[4, 1]$ | $[0, 1]$ | 128 | 512 | $10^{-3}$ | 0.2 | 0.999 | 8 |
| 1D $V$=3, $k$=3 | Shared-head | $[1, 1]$ | $[1, 1]$ | 256 | 1024 | $5\times10^{-4}$ | 0.2 | 0.95 | 8 |
| | Dedicated-heads | $[4, 1]$ | $[0, 1]$ | 256 | 1024 | $5\times10^{-4}$ | 0.2 | 0.95 | 8 |
| 2D VN $V$=2, $k$=5 | Shared-head | $[1, 1]$ | $[1, 1]$ | 1024 | 4096 | $5\times10^{-4}$ | 0.2 | 0.95 | 10 |
| | Dedicated-heads | $[8, 1]$ | $[0, 1]$ | 512 | 2048 | $10^{-3}$ | 0.2 | 0.95 | 5 |

**Optimizer and schedule.** All models are trained with AdamW ($\beta_1 = 0.9$; $\beta_2$ as listed). The learning-rate schedule is cosine decay with a linear warmup over 0.4% of total steps. Gradient norms are clipped to 1.0 at every step.

**Precision and hardware.** Training uses `bfloat16` automatic mixed precision on 1–4 GPUs; wall-clock time per run ranges from a few hours to about a day.

**Training objective.** The training loss is the average cross-entropy of the next-token prediction over all sequence positions:

$$\mathcal{L} = \frac{\sum_{b,j} m_{b,j} \left[ -\log q_\theta(y_{b,j} \mid x_{b,\le j}) \right]}{\sum_{b,j} m_{b,j}}.$$

The targets $y_{b,j}$ include both cell-state tokens ($\{0, \ldots, V-1\}$) and row/time separators; the model's softmax $q_\theta$ is taken over the full vocabulary. The loss mask $m_{b,j} \in \{0, 1\}$ is supplied by the dataset and equals 1 at every position on the training split, so no token is excluded. On the filtered evaluation splits (val-in-filter, val-out, test) the mask is 1 only at positions belonging to the prediction time steps ($t \ge T_0$) and 0 elsewhere, restricting the monitored loss to the region on which the model is scored.

**Checkpoint selection.** Deterministic checkpoints are selected on val-out sequence accuracy (maximized); stochastic checkpoints are selected on $D_{\mathrm{KL}}(q_{\text{true}} \| q_{\text{model}})$ evaluated on the val split (minimized). Early stopping uses patience 999 and min_delta 0.0005. The best checkpoint is the one where the selection metric last improved by at least min_delta.

### I.3 Evaluation protocol and metrics

All evaluation metrics are computed only over cell-value targets at scored time steps; separator targets are excluded.

**Deterministic teacher-forced (TF) cell accuracy.** At each cell position from $t = T_0$ through $t = T-1$, the model predicts the next cell value conditioned on the ground-truth history up to that point. TF cell accuracy is the fraction of correctly predicted cells (argmax of the model's output versus the ground-truth value).

**Deterministic teacher-forced (TF) sequence accuracy.** The fraction of trajectories in which every scored cell is predicted correctly, under the same teacher forcing as TF cell accuracy. A trajectory counts only if it has no error at any scored position.

**Deterministic autoregressive (AR) cell accuracy.** Starting from the ground-truth context (time steps 0 through $T_0-1$), the model predicts each subsequent cell via greedy decoding. Each predicted cell is fed back as input for subsequent predictions. AR cell accuracy is the fraction of generated cells that match the ground-truth trajectory.

**Stochastic AR cell accuracy.** The model generates a single sampled rollout from step $T_0$ onward at temperature 1 (unscaled softmax); cell accuracy is the fraction of sampled cells matching the stored ground-truth trajectory. The ground truth is one realization of a stochastic process, so a correct predictive distribution does not yield accuracy 1. The parenthesized $k$-gram baseline accuracy comes from an independent posterior-predictive sampling rollout. It matches the sampling procedure of the model's rollout, and it does not maximize accuracy (Appendix D.2). The $k$-gram predictor starts from the same $T_0$-step ground-truth prefix and samples each cell from the empirical conditional distribution given by causal $k$-gram counts with add-one smoothing. Both rollouts are scored against the same stored ground-truth trajectory.

**Stochastic AR $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{model}})$.** At each cell in the model's own autoregressive rollout, before sampling that cell's value: (1) $q_{\text{model}}$ is the softmax over the first $V$ output logits, corresponding to the cell states; (2) $q_{\text{kgram}}$ is computed from causal counts in the model-generated history up to that point, with add-one smoothing. The KL divergence is micro-averaged over all scored cells and trajectories. This $k$-gram reference is conditioned on the model's own history and is not taken from the independent $k$-gram rollout. Taking the softmax over the first $V$ logits is equivalent to taking the full-vocabulary softmax and renormalizing its cell-state probabilities. During autoregressive generation, separator tokens are inserted deterministically at their prescribed positions rather than sampled from the model.

**Stochastic TF cell accuracy.** Teacher-forced cell accuracy for stochastic rules uses argmax of the model's softmax over the first $V$ logits, compared against the ground-truth cell value. It is reported with the $k$-gram baseline in parentheses.

**Stochastic TF $D_{\mathrm{KL}}$.** Two divergences are reported under teacher forcing, both micro-averaged over scored cells. $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{model}})$ measures how closely the model's distribution matches the $k$-gram estimate when both are conditioned on the ground-truth history. $D_{\mathrm{KL}}(q_{\text{true}} \| q_{\text{model}})$ compares the model against the conditional distribution of the rule which generated the trajectory.

**Seeds and rollouts.** All evaluations use random seed 42. One AR rollout is generated per trajectory (both for the model and for the $k$-gram baseline). The reported values are micro-averages over 500k test trajectories.

### I.4 Full predictive results

Tables 13 and 14 expand the compact main-text Table 1 with additional metrics.

Table 13: Deterministic evaluation: full metrics on held-out test rules.

| Task | Arch. | TF seq acc | TF cell acc | AR cell acc |
|---|---|---|---|---|
| 1D $V{=}2,\ k{=}3$ | Shared-head | 0.9990 | 1.0000 | 0.9999 |
| | Dedicated-heads | 0.9994 | 1.0000 | 1.0000 |
| 1D $V{=}3,\ k{=}3$ | Shared-head | 0.9994 | 1.0000 | 1.0000 |
| | Dedicated-heads | 0.9989 | 1.0000 | 1.0000 |
| 2D VN $V{=}2,\ k{=}5$ | Shared-head | 1.0000 | 1.0000 | 1.0000 |
| | Dedicated-heads | 0.9989 | 1.0000 | 0.9999 |

Table 14: Stochastic evaluation: full metrics on held-out test trajectories. TF KL columns report teacher-forced divergence against the $k$-gram and true-rule baselines.

| | | TF KL | | TF cell acc | AR | |
|---|---|---|---|---|---|---|
| Task | Arch. | kg‖m | true‖m | (kgram) | cell acc (kgram) | KL kg‖m |
| 1D $V{=}2, k{=}3$ | Shared-head | 0.0020 | 0.0185 | 0.7491 (0.7499) | 0.5496 (0.5485) | 0.0019 |
| | Dedicated-heads | 0.0046 | 0.0211 | 0.7469 (0.7499) | 0.5488 (0.5485) | 0.0044 |
| 1D $V{=}3, k{=}3$ | Shared-head | 0.0026 | 0.0881 | 0.5669 (0.5674) | 0.3476 (0.3475) | 0.0026 |
| | Dedicated-heads | 0.0004 | 0.0859 | 0.5675 (0.5674) | 0.3475 (0.3475) | 0.0004 |
| 2D VN $V{=}2, k{=}5$ | Shared-head | 0.0014 | 0.0300 | 0.7345 (0.7351) | 0.5201 (0.5197) | 0.0013 |
| | Dedicated-heads | 0.0033 | 0.0319 | 0.7333 (0.7351) | 0.5191 (0.5197) | 0.0038 |

All deterministic models exceed 0.998 TF sequence accuracy. Among the stochastic metrics, the gap between $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{model}})$ and $D_{\mathrm{KL}}(q_{\text{true}} \| q_{\text{model}})$ reflects the estimation error of the $k$-gram approximation: the first measures the model against an estimate built from the same context the model sees, the second against the rule which generated it.

### I.5 Attention-alignment analysis

For each scored query cell, we measure the fraction of attention mass assigned to the positions predicted by the corresponding construction. For L1, the target set consists of the routing cells that carry the query configuration. For L2, the target set consists of the matching precedents, the earlier cells whose own parent configuration is a matching configuration for the query.

**Aggregation.** For a single query, the L1 score is computed per head as the sum of attention weights over the target positions, and the per-head scores are then averaged across heads. The L2 layer uses a single head, so the L2 score is the sum of attention weights over the target positions. The reported values are micro-averages of these per-query scores across all scored queries and trajectories.

**Trajectories.** Attention weights are extracted under teacher forcing from 20,000 test trajectories sampled uniformly without replacement from the test split.

### I.6 Attention ablation

We test whether the routing cells and matched precedents identified by the construction are functionally used by the model. For each scored query, we intervene on one attention row and measure the effect on next-token prediction. Four intervention conditions are compared against the unablated (*No ablation*) baseline: (1) *L1 routing cells*: the attention weights of the L1 head at the query's routing cells are set to zero and the row is renormalized; (2) *L1 random*: the same number of L1 attention weights are zeroed at random positions in the same time step, as a matched control; (3) *L2 matched precedents*: the L2 attention weights at earlier cells with a matching parent configuration are zeroed and the row is renormalized; (4) *L2 random*: the same number of L2 attention weights are zeroed at random non-matching completed cells. For architectures with multiple L1 heads, the L1 intervention is applied per head. Random controls are averaged over three seeds; queries for which a disjoint random set cannot be formed are left unmodified.

Interventions are applied on the same 20,000 test trajectories as the alignment analysis, and we report TF cell accuracy under each condition.

### I.7 Layerwise probe protocol

**Probe targets.** We train linear probes on frozen post-residual representations to test what information is linearly accessible at each network stage. Two targets are probed:

1. *Neighborhood configuration*: a classification target. For a scored target cell $x_{t,z}$ at serialized position $m$, we extract the representation at the preceding query position $n = m - 1$ and pair it with the target cell's parent configuration $P_{t,z}$. For a first-column target, the query position is the preceding separator. There are $V^k$ possible configurations. The probe is a linear classifier (logistic regression) and we report balanced accuracy (macro-averaged over configurations).
2. *$k$-gram predictive distribution*: a distributional target. The target distribution over the next cell value is computed from causal $k$-gram counts in the trajectory prefix with add-one smoothing. The probe is a multinomial logistic regression trained to minimize $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{probe}})$. The plotted metric is $1 - D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{probe}}) / D_{\mathrm{KL}}(q_{\text{kgram}} \| \text{Uniform}([V]))$: a score of 0 means the probe predicts no better than uniform; 1 means exact recovery of $q_{\text{kgram}}$.

**Extraction points.** Representations are extracted at five post-residual stages: after embedding, after L1 attention, after L1 MLP, after L2 attention, and after L2 MLP. Each extraction is the full residual stream after the corresponding sub-layer (after the residual connection).

**Probe model and training.** The configuration probe uses `sklearn.linear_model.LogisticRegression` with hard configuration labels and L2 regularization, solved by L-BFGS. The predictive-distribution probe uses a custom multinomial logistic regression fit with PyTorch L-BFGS, soft-label cross-entropy, and L2 regularization; its data-fit term minimizes $D_{\mathrm{KL}}(q_{\text{kgram}} \| q_{\text{probe}})$ up to a constant independent of the probe. For both probes, the regularization strength $C$ is selected on the probe validation fold from a grid search. The maximum iteration count is 500 for hyperparameter search and 2500 for the final refit; convergence is determined by the L-BFGS tolerance.

**ECA probe split (rotating 5-fold).** The ECA test rule pool contains 25 expanded rules from 9 canonical Wolfram classes. A fixed 60/20/20 split would place only 5 rules in the probe test fold, and the reported probe score would depend heavily on which 5 rules are picked. To eliminate this rule-level nuisance variability, ECA uses a rotating 5-fold split: the 25 test rules are partitioned into 5 groups of 5, and each fold uses 15 rules for training, 5 for validation, and 5 for testing. Every rule appears once as a test rule and once as a validation rule across the 5 folds. Reported scores are averaged across folds.

**Other deterministic settings (V=3, 2D VN).** The 1D $V{=}3, k{=}3$ and 2D VN test pools contain thousands of rules, so a single fixed 60/20/20 rule-level split is already well-averaged and no rotation is used.

**Stochastic probe splits.** Because each stochastic trajectory has a unique rule, the probe split is a fixed 60/20/20 split over trajectories. The rule table bytes are hashed to assign a rule ID per trajectory, and the same splitting logic as for deterministic probes is applied.

**Balancing.** For deterministic probes, cells are balanced across neighborhood configurations by per-label stratified subsampling. For stochastic probes, the target is the causal $k$-gram predictive distribution on the trajectory prefix.

### I.8 Representation visualizations with t-SNE

Figure 5 shows t-SNE embeddings of frozen residual-stream representations from the Shared-head model on deterministic 1D $V{=}2, k{=}3$ (checkpoint from Table 11: heads $[1, 1]$, mlp $[1, 1]$, $d_{\text{model}} = 128$).

**Activation collection.** Representations are extracted from 6000 test-split trajectories on the ECA test-rule pool. For each of the five post-residual stages (embedding, L1 attention, L1 MLP, L2 attention, L2 MLP), we retain 800 points per neighborhood configuration, giving $8 \times 800 = 6400$ points per stage. Within each configuration, points are stratified by target value (400 per configuration $\times$ target). Sampling uses seed 42.

**Embedding.** Each stage is embedded independently. The $(6400, 128)$ activation matrix for a stage is first reduced to 50 dimensions by PCA, then to 2 dimensions by t-SNE with perplexity 30, PCA initialization, automatic learning rate, 1000 iterations, and random seed 42; other t-SNE options are the scikit-learn defaults (Euclidean metric, Barnes–Hut, angle 0.5, early exaggeration 12).

**Figure layout.** Each of the five stages occupies one column of a $2 \times 5$ panel grid; the top row colors points by neighborhood configuration and the bottom row colors by output value. Point plotting order is shuffled with a separate seed to avoid systematic overplotting.

**Additional t-SNE plots.** We extend the main-text t-SNE to three further settings that cover the axes not shown in the body: Dedicated-heads deterministic 1D (architecture comparison, Figure 6), Shared-head stochastic 1D (stochastic regime, Figure 7), and Shared-head 2D VN (multidimensional regime, Figure 8). These t-SNE plots are intended as qualitative visualizations. Each stage is projected from a high-dimensional representation into two dimensions, so visible distances and clusters need not faithfully reflect the original representation. In particular, a lack of clear clusters does not imply that the relevant information is absent. We rely on the linear probes for quantitative evidence that the configuration and predictive distribution are decodable at each stage.

### I.9 Qualitative rollout setup for Figure 1

Figure 1 uses two checkpoints trained separately for visualization purposes, one on the deterministic 1D $V{=}2, k{=}3$ setting and one on the stochastic 1D $V{=}2, k{=}3$ setting. Both checkpoints are trained on longer trajectories and a wider grid than the main-text models in order to display long rollouts at a legible scale. Wolfram Rule 110 and the Stavskaya rule are excluded from the training rule pools of the two respective checkpoints, so the trajectories shown in Figure 1 come from rules the models have never seen. These checkpoints are used only for Figure 1: none of the predictive results in Table 1, the attention-alignment analysis, or the linear-probe experiments use them. Table 15 reports their training and evaluation configurations.

**Interpreting the stochastic panel.** The Ground Truth column displays one sampled trajectory. The Autoregressive prediction column conditions on the model's own sampled history, precluding exact per-cell reproduction even for a Bayes-optimal predictor. The Teacher-forced prediction column conditions on the ground-truth history, plotting the predictive probabilities against which the reference trajectory can be evaluated.

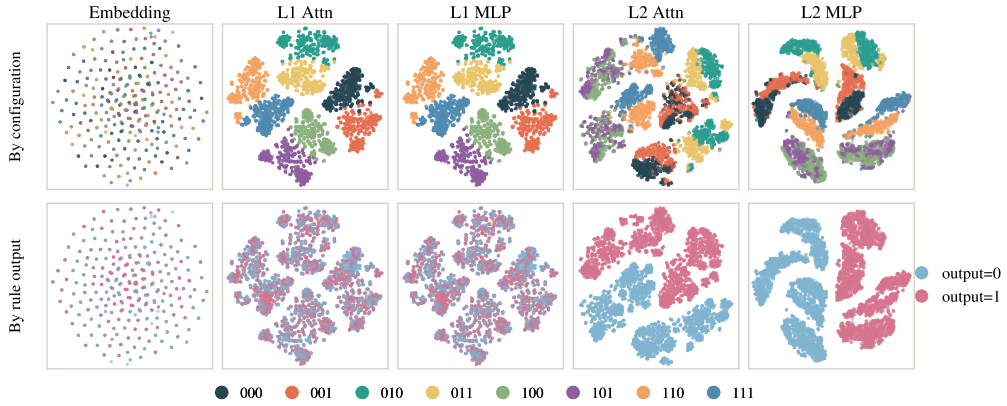


Figure 6: t-SNE of frozen residual-stream representations from the Dedicated-heads model on deterministic 1D $V{=}2, k{=}3$. Top row: colored by neighborhood configuration. Bottom row: colored by the ground-truth output value. Under the Dedicated-heads construction, the configuration should become accessible after L1 attention rather than after the L1 MLP.

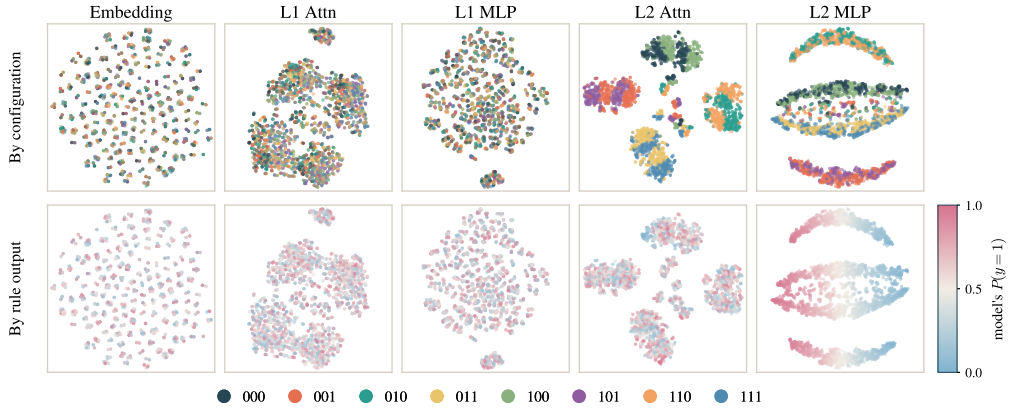


Figure 7: t-SNE of frozen representations from the Shared-head model on stochastic 1D $V{=}2, k{=}3$. Top row: colored by neighborhood configuration. Bottom row: colored by the model's predicted $P(y{=}1)$ at the corresponding query position. Because rules are drawn from a continuous Dirichlet prior, the output value at a fixed configuration is a sample from the true conditional distribution, so the model's predictive distribution varies continuously across queries with the same configuration.

Table 15: Configuration of the visualization checkpoints used for Figure 1. The deterministic panel uses greedy decoding; the stochastic panel uses temperature-1 sampling, and the displayed values are per-cell probabilities of state 1.

| | Deterministic (Rule 110) | Stochastic (Stavskaya) |
|---|---|---|
| Task | 1D $V{=}2, k{=}3$ | 1D $V{=}2, k{=}3$ |
| Training $\lvert\mathcal{G}\rvert, T, T_0$ | $64, 64, 12$ | $64, 64, 12$ |
| Plotted $\lvert\mathcal{G}\rvert, T, T_0$ | $64, 64, 12$ | $64, 64, 12$ |
| heads / mlp | $[4, 1]/[0, 1]$ | $[1, 1]/[1, 1]$ |
| $d_{\text{model}}$ / $d_{\text{ffn}}$ | $512/2048$ | $512/2048$ |
| lr, wd, epochs | $5{\times}10^{-4}$, 0.2, 10 | $10^{-3}$, 0.2, 8 |
| Batch (gbs), hardware | 1024, 2×A6000 DDP | 1024, 2×B300 DDP |
| Checkpoint selection | val-out seq acc (max) | $D_{\text{KL}}(q_{\text{true}} \Vert q_{\text{model}})$ (min) |
| Decoding at plotted step | greedy argmax | Bernoulli sampling at temperature 1 |

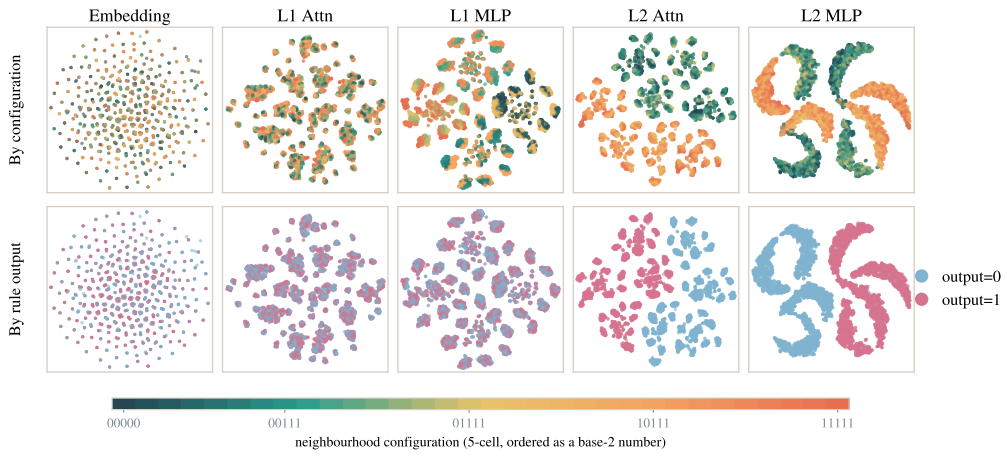


Figure 8: t-SNE of frozen representations from the Shared-head model on deterministic 2D von Neumann $V$=2, $k$=5. Top row: colored by neighborhood configuration. Bottom row: colored by the ground-truth output value. The larger neighborhood ($k = 5$, $V^k = 32$ configurations) is harder to display in two dimensions than the ECA case; per-configuration clusters are correspondingly less separated.