UNIVERSALITY AND GENERALIZATION OF CAUSAL TRANSFORMERS ACROSS CONTEXT LENGTHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long contexts are central to modern transformer systems, but most expressivity results choose a different network for each fixed sequence length. We study whether one masked transformer can approximate causal token-to-token maps uniformly over sequences of arbitrary length sampling a fixed normalized horizon. To relate sampling resolutions, we model tokens by α-Hölder sequences or, more generally, a common modulus of continuity. Our notion of continuity across resolutions characterizes the causal families admitting uniform approximation on these compact input classes by a single transformer with length-independent parameters. The result extends to the infinite-length mean-field limit, where tokens form continuous curves and masked attention becomes a causal time integral. For bounded regression with target maps satisfying a β-smooth stability condition defined using regular test functions, quantitative approximation yields a generalization bound: exact empirical risk minimization over suitably sized bounded-weight transformers gives root mean-square prediction error $O((\log\log N/\log N)^{\beta/(d+2)})$ from N iid labeled sequences. The bound holds at fixed confidence on the same sampling distribution, with d the token dimension and no maximum-length factor. Finally, experiments on physical time series support the Hölder-regular token model at observed scales, with dataset-dependent fitted exponents, whereas text input embeddings provide a contrasting case. Native and dense sampling, shuffled controls, and refinement checks delimit this empirical regularity regime.

1 INTRODUCTION

Long-context transformers apply one learned rule across sequence lengths, yet most universality results fix the length before choosing the network. We show that common input regularity, together with asymptotic continuity of the target across resolutions, enables uniform causal approximation at every resolution, connects the discrete model to its continuous-time limit, and yields approximation and statistical learning rates under additional target regularity. We also clarify the role of positional information and examine the input geometry empirically.

Transformers and long-context architectures. The original transformer combines multi-head attention, pointwise feed-forward layers, and positional encodings (Vaswani et al., 2017). Long-context variants use recurrence, sparse or efficient attention, and position extrapolation (Dai et al., 2019; Beltagy et al., 2020; Dao et al., 2022; Ding et al., 2024), while RoPE, ALiBi, and positional interpolation target generalization beyond the training window (Su et al., 2024; Press et al., 2022; Chen et al., 2023). Short-to-long generalization has also been studied under an idealized learning rule (Huang et al., 2025). We ask a different expressivity question: can one parameter set approximate a causal rule uniformly over every resolution? Here, "all resolutions" means increasingly fine samples of a fixed normalized horizon under a common regularity model, not unrestricted growth of a symbolic string. In infinite-token limits, unmasked attention naturally acts on empirical token distributions (Vuckovic et al., 2020); causal masking instead requires token–position information to preserve chronology (Castin et al., 2024; Furuya et al., 2025). We compactify all grids jointly and recover temporal attention as the refinement limit of the same finite architecture. This token-number limit is distinct from a continuous-depth limit (Karagodin et al., 2024).

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.

Neural-network and transformer universality. Classical results establish universality of feed-forward networks on compact finite-dimensional sets and of neural networks between function spaces (Cybenko, 1989; Leshno et al., 1993; Chen & Chen, 1995). Neural operators extend this viewpoint to discretization-invariant maps, without generally enforcing causality (Kovachki et al., 2023). Transformer universality is known for fixed-length equivariant or position-aware maps and for sparse or shallow variants (Yun et al., 2020a;b; Kajitsuka & Sato, 2024); related results treat fixed-window causal models, shift-equivariant infinite sequences, and structured next-token targets (Yang et al., 2023; Takakura & Suzuki, 2023; Sander & Peyré, 2025). Quantitative results cover fixed-length smooth targets and bidirectional sliding-window maps under distributional error (Jiang & Li, 2024; Takakura & Suzuki, 2023). Our result instead controls one causal architecture uniformly over all resolutions and its continuous-time limit. For $\beta > 1$, our β-smooth target condition is stability in a smooth-test metric on token–position laws, rather than generic Fréchet C^β regularity. Logarithmic statistical rates also arise in functional regression (Meister, 2016); neural-operator analyses connect approximation complexity to empirical-risk guarantees (Liu et al., 2024; Kovachki et al., 2024). Our additional step controls the capacity of bounded-weight causal transformers uniformly in context length. Positional features remove permutation obstructions in fixed-length universality (Yun et al., 2020a), although some relative schemes remain limited (Luo et al., 2022). In our all-resolution setting, normalized position also records chronology and grid resolution. Appendix A proves a sharper obstruction: even when total length is known and position enters arbitrary normalized attention scores, the normalized clock cannot be recovered if position is absent from values, residual features, and pointwise maps.

Universality for causal sequence operators. Before transformers, recurrent networks and causal convolutions encoded non-anticipativity (Hochreiter & Schmidhuber, 1997; Hanson & Raginsky, 2019). Their approximation theory either fixes the sequence length or horizon (Funahashi & Nakamura, 1993; Song et al., 2023), or obtains time-uniform guarantees through fading memory (Grigoryeva & Ortega, 2018; Gonon & Ortega, 2021). Rates are available for temporal convolutions and random reservoirs (Hanson & Raginsky, 2019; Gonon et al., 2023), while recent causal neural-operator and temporal-integral constructions treat continuous paths on prescribed spans (Acciaio et al., 2024; Galimberti et al., 2026; Cuchiero et al., 2026). Our normalized-horizon setting may retain the entire prefix: compactness comes from regularity across sampling resolutions, and one standard masked-attention architecture provides uniform approximation and, under additional target regularity, parameter rates.

Contributions and organization. Section 2 introduces an all-resolution causal setting in which token grids share an arbitrary admissible modulus and finite targets need only converge asymptotically, rather than obey exact projective identities, to a continuous path operator. Our main result, Theorem 2.4, shows that one shallow masked transformer—one attention block and one shared pointwise ReLU network—approximates every target in such a family uniformly over sequence length, input, and prediction index; its parameters may depend on the target and accuracy, but not on length. The converse characterizes their uniform closure (Corollary D.13). The completed-space proof gives, with the same parameters, temporal-attention universality for the continuous path limit (Theorem 2.6). For normalized β-smooth teachers, Theorem 2.9 gives root mean-square prediction error $O((\log\log N/\log N)^{\beta/(d+2)})$ at fixed confidence from N iid variable-length sequences, using clipped exact empirical risk minimization over suitably sized bounded-weight transformers. This same-distribution bound relies on quantitative uniform approximation (Appendix B) and a capacity estimate independent of context length (Appendix G). Finally, Section 3 finds dataset-dependent finite-scale increment decay in physical signals and continuous pretrained content features, while text input embeddings provide a contrasting case. Native and dense patch sampling, shuffled controls, and refinement diagnostics delimit this empirical support for the Hölder input model; the detailed proofs and protocol are collected in the appendices.

2 All-resolution causal universality

We ask whether one causal transformer can approximate a family of prediction rules uniformly across all context lengths. A shared architecture, a common input modulus, and asymptotic cross-resolution compatibility make this possible. We first establish discrete and continuous-time universality, then derive a learning guarantee under additional target regularity.

Fix dimensions $d, d' \geq 1$ and $[n] := \{1, \ldots, n\}$. Write $z = (z_i)_{i=1}^n$ for discrete inputs and $x = (x(t))_{t\in[0,1]}$ for continuous inputs. We use $\|\cdot\|$ for Euclidean norms, $|\cdot|$ for absolute values, and $\|\cdot\|_\infty$ for uniform or maximum norms.

2.1 Discrete causal transformer

The architecture is defined once on the space of all nonempty finite sequences. Its dimensions, weights, and biases are independent of sequence length, so the same formulas apply for every $n \geq 1$.

A common domain across lengths. For each $p \geq 1$, the disjoint union

$$\mathsf{Seq}_p := \bigsqcup_{n\geq 1} (\mathbb{R}^p)^n \tag{1}$$

collects all nonempty finite sequences of p-dimensional vectors.

Residual masked multi-head attention. Hidden states $u_i \in \mathbb{R}^p$ are internal representations of input tokens $z_i \in \mathbb{R}^d$. Each head summarizes the current prefix; the residual connection adds u_i. For $u \in \mathsf{Seq}_p$ of length n and $i \in [n]$, define

$$\mathrm{MAtt}_\theta(u)_i := u_i + \sum_{r=1}^{H} W_r \frac{\sum_{j=1}^{i} e^{\langle Q_r u_i, K_r u_j\rangle} V_r u_j}{\sum_{j=1}^{i} e^{\langle Q_r u_i, K_r u_j\rangle}}. \tag{2}$$

Equation (2) therefore defines a single operator $\mathrm{MAtt}_\theta : \mathsf{Seq}_p \to \mathsf{Seq}_p$. On each component, the input length determines the summation range, while the expression and parameters remain unchanged. For head r, if the query–key dimension is q_r and the value dimension is v_r, the matrices in (2) have dimensions

$$Q_r, K_r \in \mathbb{R}^{q_r\times p}, \qquad V_r \in \mathbb{R}^{v_r\times p}, \qquad W_r \in \mathbb{R}^{p\times v_r}. \tag{3}$$

Here $\theta := ((Q_r, K_r, V_r, W_r))_{r=1}^H$. Thus the inner product in the score is taken in $\mathbb{R}^{q_r}$, each normalized head value lies in $\mathbb{R}^{v_r}$, and the output again lies in $\mathbb{R}^p$.

We use ∞ to distinguish continuous-time realizations. Along any grid refinement whose hidden-state interpolants converge uniformly, masked softmax averages converge to causal temporal integrals. For a continuous hidden path $u : [0, 1] \to \mathbb{R}^p$, the limiting attention block is

$$\mathrm{MAtt}^\infty_\theta(u)(t) := u(t) + \sum_{r=1}^{H} W_r \begin{cases} \dfrac{\int_0^t e^{\langle Q_r u(t), K_r u(s)\rangle} V_r u(s)\, \mathrm{d}s}{\int_0^t e^{\langle Q_r u(t), K_r u(s)\rangle}\, \mathrm{d}s}, & t > 0, \\ V_r u(0), & t = 0. \end{cases} \tag{4}$$

Shared pointwise networks. A shared tokenwise affine layer applies one affine map at every token and every length:

$$G_{A,b} : \mathsf{Seq}_p \to \mathsf{Seq}_q, \qquad (G_{A,b}(u))_i := Au_i + b, \qquad A \in \mathbb{R}^{q\times p}, \quad b \in \mathbb{R}^q.$$

Its continuous-time realization applies the same affine map at every time: $G_{A,b,\infty}(u)(t) := Au(t) + b$. A shared ReLU network applies the same finite-dimensional map to every token. For input dimension p, output dimension q, and hidden width $M \geq 1$, a one-hidden-layer pointwise ReLU network is the map

$$\mathrm{MLP}_\eta(u) := A_2\rho(A_1u + b_1) + b_2, \qquad u \in \mathbb{R}^p, \tag{5}$$

where ρ is the ReLU activation function acting coordinatewise and

$$A_1 \in \mathbb{R}^{M\times p}, \quad b_1 \in \mathbb{R}^M, \qquad A_2 \in \mathbb{R}^{q\times M}, \quad b_2 \in \mathbb{R}^q; \qquad \eta := (A_1, b_1, A_2, b_2). \tag{6}$$

We use the same notation for its pointwise action on sequences: $\mathrm{MLP}_\eta : \mathsf{Seq}_p \to \mathsf{Seq}_q$, with $(\mathrm{MLP}_\eta(u))_i := \mathrm{MLP}_\eta(u_i)$.

A finite causal transformer is any finite composition of these masked-attention blocks, shared affine layers, and pointwise ReLU networks with compatible widths. Its finite parameter list is shared across lengths.

The shallow universal architecture. On regular input classes and continuously extendable target families introduced below, one masked-attention block followed by one pointwise MLP suffices for qualitative universality when width may grow with the target and accuracy. Given an integer $H \geq 1$, we set $p_H := d + H + 2$ and define the fixed tokenwise lift, $\mathcal{E}_H : \mathsf{Seq}_{d+1} \to \mathsf{Seq}_{p_H}$, for $y = (y_i)_{i=1}^n \in (\mathbb{R}^{d+1})^n$ by

$$\mathcal{E}_H(y) := \big((y_i, 1, 0_H)\big)_{i=1}^n \in (\mathbb{R}^{p_H})^n. \tag{7}$$

The attention input is $u_i := \mathcal{E}_H(y)_i$. The lift contains no learned parameters and acts identically at every token.

All H attention heads are taken to be scalar ($q_r = v_r = 1$), so

$$Q_r, K_r, V_r \in \mathbb{R}^{1 \times p_H}, \qquad W_r \in \mathbb{R}^{p_H \times 1}, \qquad r \in [H]. \tag{8}$$

At each token, the final H coordinates are "work" coordinates. In the construction, the W_r's have disjoint ranges, so head r stores one scalar probe in work coordinate r. For an MLP hidden width $M \geq 1$ and output dimension $q = d'$ in (6), we define

$$T_\Theta := \mathrm{MLP}_\eta \circ \mathrm{MAtt}_\theta \circ \mathcal{E}_H, \qquad \Theta := (\theta, \eta). \tag{9}$$

This defines $T_\Theta : \mathsf{Seq}_{d+1} \to \mathsf{Seq}_{d'}$ by the same composition at every length. In particular, $A_1 \in \mathbb{R}^{M \times p_H}$ and $A_2 \in \mathbb{R}^{d' \times M}$. The integers H, M, all matrices, and all biases are fixed across sequence lengths, but may depend on the target and accuracy.

Normalized position in the hidden state. We append normalized position i/n to each input token, making it available through the residual connection. For $z \in (\mathbb{R}^d)^n$, define the positional lift

$$\phi_n(z)_i := (z_i, i/n) \in \mathbb{R}^{d+1}. \tag{10}$$

The transformer receives $y := \phi_n(z)$. No separate length channel is needed: one uniform-attention head computes the prefix mean position, and the residual position gives $2i^{-1} \sum_{j=1}^{i} (j/n) - i/n = 1/n$. Thus the resolution is accessible even though the current position alone does not identify it. Appendix A shows why position confined to normalized attention scores cannot replace this residual information.

2.2 Discrete causal targets with a continuous path extension

Uniform universality across arbitrary context lengths rests on two joint hypotheses on the context and the target: a common modulus places all token grids in one compact all-resolution family, and the discrete causal targets converge to a single continuous path operator under grid refinement.

Regular token sequences of arbitrary length. A common concave modulus controls every resolution and is preserved exactly by piecewise-affine interpolation.

Definition 2.1 (Admissible common modulus)**.** An *admissible common modulus* is a continuous, nondecreasing, concave function $\omega : [0, 1] \to [0, \infty)$ satisfying $\omega(0) = 0$.

In particular, $\omega(r) \to 0$ as $r \downarrow 0$. The zero modulus is allowed and gives constant token paths, while $\omega_{\alpha,L}(r) := Lr^\alpha$ is admissible for every $0 < \alpha \leq 1$ and $L > 0$.

Let $\Omega := \overline{B_{\mathbb{R}^d}(0, 1)}$, and fix an admissible common modulus ω. For each length n, we define

$$X_n^\omega := \left\{ z = (z_1, \ldots, z_n) \in \Omega^n : \|z_j - z_i\| \leq \omega\left(\frac{|j - i|}{n}\right), \quad i, j \in [n] \right\}. \tag{11}$$

The estimate is imposed on every pair, not only on adjacent tokens. This distinction is essential for a general modulus: adjacent control gives only $\|z_j - z_i\| \leq |j - i|\, \omega(1/n)$, which need not imply the required $\omega(|j - i|/n)$ estimate. For the linear modulus $\omega(r) = Lr$, the all-pairs and adjacent conditions are equivalent. This models regular input vectors, not arbitrary sequences of token embeddings.

Continuous path completion. Sampling and interpolation identify the discrete token grids with approximations of a single compact class of continuous paths. Define

$$X_\infty^\omega := \{x \in C([0,1];\Omega) : \|x(t) - x(s)\| \le \omega(|t-s|), \quad s,t \in [0,1]\}. \tag{12}$$

For a path $x : [0,1] \to \mathbb{R}^d$, define its samples by $\mathcal{S}_n x := (x(j/n))_{j=1}^n$. For a sequence $z \in (\mathbb{R}^d)^n$, let the interpolant $\mathcal{I}_n z$ be constant and equal to z_1 on $[0, 1/n]$, affine between consecutive grid values, and satisfy

$$(\mathcal{I}_n z)(i/n) = z_i, \qquad i \in [n]. \tag{13}$$

The same notation applies to output interpolation. Lemma D.1 proves, for every $x \in X_\infty^\omega$,

$$\mathcal{I}_n X_n^\omega \subset X_\infty^\omega, \qquad \mathcal{S}_n(X_\infty^\omega) = X_n^\omega, \qquad \|\mathcal{I}_n \mathcal{S}_n x - x\|_\infty \le \omega(1/n). \tag{14}$$

Continuously extendable causal targets. The finite target maps need not agree exactly across context lengths; we require only that they be causal at each length and converge uniformly to one continuous path operator.

Definition 2.2 (Continuously extendable discrete causal family)**.** A discrete family $(F_n^\star)_{n\ge1}$ is a *continuously extendable causal family* if each map $F_n^\star : X_n^\omega \to (\mathbb{R}^{d'})^n$ is continuous and prefix-causal, meaning that, for all $z, z' \in X_n^\omega$ and $i \in [n]$, $z_{1:i} = z'_{1:i}$ implies $F_n^\star(z)_i = F_n^\star(z')_i$, and there exists a continuous map $F_\infty^\star : X_\infty^\omega \to C([0,1];\mathbb{R}^{d'})$ such that

$$\sup_{z \in X_n^\omega} \|\mathcal{I}_n(F_n^\star(z)) - F_\infty^\star(\mathcal{I}_n z)\|_\infty \longrightarrow 0 \qquad (n \to \infty). \tag{15}$$

Such an $F_\infty^\star$ is a *continuous path extension* of the discrete family.

Controlled dynamical systems provide a concrete example.

Example 2.3 (Controlled ODEs)**.** Fix $y_0 \in \mathbb{R}^{d'}$ and a continuous vector field $b : \mathbb{R}^{d'} \times \Omega \to \mathbb{R}^{d'}$, globally Lipschitz in its state argument uniformly over controls. For $x \in X_\infty^\omega$, let y_x solve

$$\dot{y}_x(t) = b(y_x(t), x(t)), \quad y_x(0) = y_0, \qquad F_\infty^\star(x) := y_x, \quad F_n^\star(z)_i := y_{\mathcal{I}_n z}(i/n).$$

Thus the discrete map samples the exact state trajectory driven by the reconstructed tokens. This family satisfies Definition 2.2: the extension error in (15) is at most M/n, with M depending only on b, y_0. Appendix E gives the construction, uniform estimates, and proof.

Although causality is imposed only at finite resolutions, it passes to the extension: $x|_{[0,t]} = x'|_{[0,t]}$ implies $F_\infty^\star(x)(t) = F_\infty^\star(x')(t)$. Lemma D.4 proves this fact and also shows that the extension is unique.

Compatibility with transformers across resolutions. For every fixed parameter list Θ, the family induced by the transformer in (9) and the normalized positional lifts, $\mathcal{T}_{\Theta,n}(z) := T_\Theta(\phi_n(z))$, $n \ge 1$, is continuously extendable and causal in the sense of Definition 2.2. Corollary D.8 proves this for every fixed finite causal transformer.

Cross-resolution compatibility is asymptotic. No separate output modulus is required: the compact image of the continuous extension supplies the needed equicontinuity. We write $\mathbf{F}^\star := ((F_n^\star)_{n\ge1}, F_\infty^\star)$ for the completed family. Lemma D.4 shows that (15) is equivalent to $\mathcal{I}_{n_k}(F_{n_k}^\star(z^k)) \to F_\infty^\star(x)$ uniformly whenever $n_k \to \infty$ and $\mathcal{I}_{n_k} z^k \to x$ uniformly. It allows arbitrary continuous causal changes at finitely many lengths and imposes no exact identity either with $F_\infty^\star$ or between two finite lengths. Such projective consistency would be unnatural here: index i represents time i/n, so even the clock $F_n^\star(z)_i := i/n$ violates a same-index identity across lengths.

2.3 Universality theorems

We establish universality for the shared discrete architecture and its continuous-time realization. Section 2.4 then turns target regularity into a statistical learning bound.

Universality for discrete tokens. The first result is stated entirely at finite resolutions. Continuous paths appear only through the extension hypothesis; neither temporal attention nor the measure representation used in the proof enters the conclusion.

Theorem 2.4 (All-resolution universality for discrete tokens)**.** *Fix an admissible common modulus ω. Let $(F_n^\star)_{n\geq 1}$ be a continuously extendable causal family on $(X_n^\omega)_{n\geq 1}$ in the sense of Definition 2.2. For every $\varepsilon > 0$, there exist integers $H, M \geq 1$ and a parameter list $\Theta = (\theta, \eta)$, independent of n, with residual width $p_H = d + H + 2$, scalar-head dimensions $Q_r, K_r, V_r \in \mathbb{R}^{1\times p_H}$, $W_r \in \mathbb{R}^{p_H\times 1}$, and readout dimensions $A_1 \in \mathbb{R}^{M\times p_H}$, $b_1 \in \mathbb{R}^M$, $A_2 \in \mathbb{R}^{d'\times M}$, $b_2 \in \mathbb{R}^{d'}$, such that the transformer T_Θ defined by (7)-(9) satisfies*

$$\sup_{n\geq 1}\sup_{z\in X_n^\omega}\max_{i\in[n]}\|T_\Theta(\phi_n(z))_i - F_n^\star(z)_i\| < \varepsilon. \tag{16}$$

Every attention matrix Q_r, K_r, V_r, W_r can be chosen with Euclidean-induced operator (spectral) norm at most one.

Sketch of proof. The common modulus compactifies finite evaluation states by adding continuous-path limits. Joint token–position prefix laws retain exactly the causal information; causality and compatibility therefore make the target a continuous function of these laws. Scalar attention probes are log-Laplace derivatives and separate distinct laws. Stone–Weierstrass approximates the target by a polynomial of finitely many probes: one attention block computes them in parallel, and a shared ReLU readout approximates the polynomial. Compactness makes the error uniform over all lengths. Appendix D.1 gives the detailed reading guide; Proposition D.5 and Lemmas D.9–D.10 supply the main ingredients. □

The converse also holds: Corollary D.13 proves that continuous extendability is necessary for uniform all-resolution approximation, even by arbitrary finite-depth transformers. Thus shallow and finite-depth architectures have the same uniform closure. Each approximating network is fixed across lengths; width, readout weights, and, in the larger class, depth may vary with accuracy. The construction keeps attention scores uniformly bounded, so increasingly sharp attention is unnecessary. However, position cannot be confined to scores: Appendix A proves an error of at least $1/2$ on the normalized clock in this case, even if total length is known. These are expressivity guarantees, not guarantees of learning or length-independent computation.

Remark 2.5 (Width versus depth). The parallel construction uses H scalar heads and residual width $p_H = d + H + 2$ to store H probes. Alternatively, the Stone–Weierstrass polynomial can be evaluated sequentially, reusing coordinates for the current probe, running monomial product, and accumulated sum. Lemma D.11, following Furuya et al. (2025), gives residual width $d + 1 + 3d'$ and at most d' scalar heads per block. Depth and pointwise hidden widths may then depend on the target and accuracy, whereas residual width and head count remain fixed; none depends on sequence length. Appendix B analyzes only the shallow parallel construction.

Continuous-time limit and universality for paths. Replacing masked sums by the temporal integrals in (4) defines the same architecture on paths. Set $\phi_\infty(x)(t) := (x(t), t)$ and $T_{\Theta,\infty} := \mathrm{MLP}_\eta \circ \mathrm{MAtt}_\theta^\infty \circ \mathcal{E}_H$, where $(\mathcal{E}_H y)(t) := (y(t), 1, 0_H)$ and the readout acts pointwise.

Theorem 2.6 (Continuous-time causal universality)**.** *Fix an admissible common modulus ω. Let $F_\infty^\star : X_\infty^\omega \to C([0,1];\mathbb{R}^{d'})$ be continuous in the uniform norm and causal: $x|_{[0,t]} = x'|_{[0,t]}$ implies $F_\infty^\star(x)(t) = F_\infty^\star(x')(t)$. For every $\varepsilon > 0$, there exist integers $H, M \geq 1$ and parameters Θ, with the dimensions and attention-matrix norm bounds of Theorem 2.4, such that*

$$\sup_{x\in X_\infty^\omega}\|T_{\Theta,\infty}(\phi_\infty(x)) - F_\infty^\star(x)\|_\infty < \varepsilon. \tag{17}$$

If $F_\infty^\star$ extends a discrete family as in Definition 2.2, the same parameters can also satisfy (16).

Thus a continuous causal operator $F_\infty^\star$ can be approximated without specifying any discrete target family. Temporal attention is also the refinement limit: for fixed Θ, if $n_k \to \infty$ and $\mathcal{I}_{n_k} z^k \to x$ uniformly with $z^k \in X_{n_k}^\omega$, then $\mathcal{I}_{n_k} T_\Theta(\phi_{n_k}(z^k)) \to T_{\Theta,\infty}(\phi_\infty(x))$ uniformly. Appendix F details the realization and proves this convergence (Proposition F.2); the common universality proof is in Appendix D.

2.4 Quantitative rates and generalization bound

To obtain approximation and learning rates uniform over context lengths, we impose β-smooth regularity on the common discrete–path target (Appendix G.5).

Quantitative rates for β-smooth targets. We impose regularity after factoring the targets through measures of causal histories (Appendix B.1). For $E := \Omega \times [0,1]$, let $\mathcal{M}_\omega \subset \mathcal{P}(E)$ be the weak closure of the prefix laws $\mu_{n,z,i} := i^{-1}\sum_{j\le i}\delta_{(z_j,j/n)}$, for $z \in X_n^\omega$. The endpoint map $\mathfrak{e}(\mu_{n,z,i}) := (z_i, i/n)$ extends continuously to $\mathcal{M}_\omega$. Continuously extendable families factor as $F_n^\star(z)_i = f^\star(\mu_{n,z,i})$, with a unique continuous $f^\star : \mathcal{M}_\omega \to \mathbb{R}^{d'}$ (Proposition D.5). For $\beta \ge 1$, the law distance below is a Hölder integral probability metric, closely related to smooth Wasserstein metrics (Block et al., 2022; Gaunt & Li, 2023).

Definition 2.7 (β-smooth regularity class)**.** For $\beta > 0$, put $\beta_- := \min\{\beta, 1\}$, $\beta_+ := \max\{\beta, 1\}$, and $\delta_e(\mu,\nu) := \|\mathfrak{e}(\mu) - \mathfrak{e}(\nu)\|_\infty$. Define on $\mathcal{M}_\omega$, with smoothness index β,

$$W_{1,\beta}(\mu,\nu) := \Big[\sup_{\|\varphi\|_{C^{\beta_+}(E)}\le 1}\int_E \varphi\,\mathrm{d}(\mu-\nu)\Big]^{\beta_-}, \qquad \Delta_\beta := W_{1,\beta} + \delta_e^{\beta_-}. \tag{18}$$

A continuously extendable family $\mathbf{F}^\star$ is *β-smooth* if $f^\star$ is Lipschitz on $(\mathcal{M}_\omega, \Delta_\beta)$. Write $R_{\beta;\omega}(\mathbf{F}^\star)$ for its optimal Lipschitz constant. The normalized class $\mathcal{F}^\beta(\omega)$ is the set of such families $\mathbf{F}^\star$ with $R_{\beta;\omega}(\mathbf{F}^\star) \le 1$ and $\|f^\star\|_\infty \le 1$ on $\mathcal{M}_\omega$. As defined in Appendix B.1, $C^\sigma(E)$ uses the full isotropic restriction norm from $[-1,1]^d \times [0,1]$, with $C^m = C^{m-1,1}$ at integer orders. At $\beta = 1$, $W_{1,1}$ is equivalent to Wasserstein-1 (equal with Lipschitz-seminorm normalization; Lemma B.1).

Example 2.8 (Regular controlled integrators)**.** Define $F_\infty^\star(x) := y_x$ as in Example 2.3 with $b(y,u) = g(u)$, where $g \in C^\beta(\Omega;\mathbb{R}^{d'})$. The targets $F_n^\star(z)_i := y_0 + n^{-1}\sum_{j=1}^i g(z_j)$ use the constant rate $g(z_j)$ on each interval $((j-1)/n, j/n]$. These right-endpoint Riemann sums extend to the same continuous map. Their history factor $f^\star(\mu) = y_0 + t\int g(u)\,\mathrm{d}\mu(u,s)$, with $t := \mathfrak{e}(\mu)_{d+1}$, satisfies $R_{\beta;\omega}(\mathbf{F}^\star) \le C_{d,d',\beta}\,\|g\|_{C^\beta}$, with an ω-independent upper bound. Unlike Example 2.3, these maps do not integrate the piecewise-affine control exactly; Appendix E.3 details the distinction and proves the bound.

Our regularity condition is Hölder stability for $0 < \beta \le 1$, and stability through smooth token–time averages with Lipschitz endpoint dependence for $\beta > 1$, not generic Fréchet smoothness (Remark B.6). Theorem B.3 (Appendix B) gives uniform discrete–path error $O(R_{\beta;\omega}(\log\log p/\log p)^{\beta/(d+2)})$ with p parameters. For Hölder inputs, finite-description lower bounds are also logarithmically slow (Appendix B.5); matching approximation and generalization lower bounds remain open.

Generalization bound. We fit the target family by squared-loss empirical risk minimization (ERM). For H scalar heads and an M-unit readout, counting all matrix and bias entries, including zeros, gives $\mathrm{par}(T_\Theta) := 4Hp_H + (p_H + d' + 1)M + d'$. The lift is parameter-free. For a deterministic integer budget p, let $\mathfrak{T}_p$ contain all shallow transformers (9), over integer $H, M \ge 1$, with

$$\mathrm{par}(T_\Theta) \le p, \qquad \|\theta\|_\infty \le 1, \qquad \|\eta\|_\infty \le p+1, \tag{19}$$

where the parameter norms bound every matrix entry and bias in their respective blocks. For training sequences $(n_k, Z^k, Y^k)_{k=1}^N$, with $Z^k \in X_{n_k}^\omega$ and $Y^k \in (\mathbb{R}^{d'})^{n_k}$, use the Frobenius norm over tokens and define

$$\widehat{\mathcal{L}}_N(T_\Theta) := \frac{1}{N}\sum_{k=1}^N \frac{\|T_\Theta(\phi_{n_k}(Z^k)) - Y^k\|_F^2}{n_k}, \quad \widehat{\mathcal{L}}_N(T_{\widehat\Theta_N}) \le \inf_{T_\Theta\in\mathfrak{T}_p}\widehat{\mathcal{L}}_N(T_\Theta) + \tau_{\mathrm{opt}}. \tag{20}$$

Here $\widehat\Theta_N$ measurably selects a network in $\mathfrak{T}_p$ from the data, within $\tau_{\mathrm{opt}} \ge 0$ of the minimum empirical loss (zero for exact ERM). We do not bound this optimization gap for SGD (Appendix G). Define the predicted family $\widehat F := (\widehat F_n)_{n\ge1}$ by $\widehat F_n(z)_i := \mathrm{clip}(T_{\widehat\Theta_N}(\phi_n(z))_i)$, where $\mathrm{clip}(v) := v/\max\{1, \|v\|\}$ projects onto the output unit ball, giving a budget-independent loss bound for concentration with bounded labels. Conditional on the training data, the prediction risk on an independent test sequence from the same law is

$$\mathcal{R}_{\mathrm{pred}}(\widehat F) := \mathbb{E}_{(n,Z)}\left[\frac{1}{n}\sum_{i=1}^n \left\|\widehat F_n(Z)_i - F_n^\star(Z)_i\right\|^2\right]. \tag{21}$$

Theorem 2.9 (Generalization from variable-length sequences)**.** *Fix an admissible common modulus ω, $\beta > 0$, and a teacher $\mathbf{F}^\star \in \mathcal{F}^\beta(\omega)$. Let $(n_k, Z^k, Y^k)_{k=1}^N$ be iid copies of (n, Z, Y), with $Z \in X_n^\omega$, $\|Y_i\| \le 1$, and $\mathbb{E}[Y_i \mid n, Z_{1:i}] = F_n^\star(Z)_i$, for each i almost surely on $\{n \ge i\}$. There are constants $C \ge 1$ and $p_0 \ge 16$, depending only on (d, d', β), such that for every deterministic integer $p \ge p_0$, $N \ge 2$, and $0 < \delta < 1$, with probability at least $1 - \delta$, every estimator satisfying (20) obeys*

$$\mathcal{R}_{\mathrm{pred}}(\widehat{F}) \le C \left[\left(\frac{\log\log p}{\log p} \right)^{2\beta/(d+2)} + \frac{p \log(CpN) + \log(2/\delta)}{N} \right] + 2\tau_{\mathrm{opt}}. \tag{22}$$

The constants are independent of the input modulus, the sampling distribution, and the context lengths; no upper bound or moment assumption on n is needed.

Sketch of proof. Corollary B.5 supplies a bounded-weight comparator with the stated approximation error. Normalized-attention stability gives the clipped class a uniform cover of logarithmic size $O(p \log(p/\epsilon))$, independent of context length (Lemma G.2). Sequence-averaged excess losses have variance controlled by prediction risk, so Bernstein concentration gives the N^{-1} statistical term. Projection decreases empirical loss; comparing the clipped raw ERM with the comparator then adds $2\tau_{\mathrm{opt}}$. Appendix G gives the proof. □

The bound separates approximation from estimation. For exact ERM, $p := \lfloor \sqrt{N} \rfloor$ gives root mean-square rate $O((\log\log N / \log N)^{\beta/(d+2)})$ at fixed confidence and large N. Here N counts independent sequences, allowing dependent tokens.

3 Empirical diagnostics of Hölder path geometry

The common-modulus assumption links sequences sampled at different resolutions of a fixed horizon. We examine its Hölder specialization through the decay of worst increments in measured signals and pretrained representations. The experiments reveal a positive but representation-dependent finite-scale regime; they do not certify one common Hölder ball at all resolutions.

A common scale diagnostic. Classical multiscale methods assess roughness through the log–log scaling of increments or wavelet coefficients (Gneiting et al., 2012; Abry et al., 2015). We use a maximum over positions, rather than an average, to probe the worst-case increments controlled by our uniform Hölder hypothesis. For vectors $z_0, \dots, z_{N-1} \in \mathbb{R}^d$ on a uniform time grid of spacing Δ, define, for $1 \le r < N$ and $h_r := r\Delta$,

$$M_\infty(r) := \max_{0 \le k < N-r} d^{-1/2} \|z_{k+r} - z_k\|_2 . \tag{23}$$

We evaluate dyadic lags with $h_r \le 1/4$ to limit finite-window effects, and fit each window on the finer band:

$$\log M_\infty(r) = \log C + \widehat{\alpha}_\infty \log h_r + \varepsilon_r, \qquad h_r \le 1/32. \tag{24}$$

We use $\Delta := 1/(N-1)$, or $\Delta := s/(N_{\mathrm{raw}} - 1)$ for patches extracted every s raw measurements to preserve physical time. We report mean slopes and sample SD across windows, require at least four scales per primary fit, and retain all declared alternative bands. An L-Hölder-α path satisfies $M_\infty(r) \le L h_r^\alpha / \sqrt{d}$; a fitted slope alone does not establish this bound. Thus C is a fitted prefactor and $\widehat{\alpha}_\infty$ a finite-scale diagnostic, not a certified asymptotic uniform exponent.

Physical signals and continuous content. Physical recordings admit fixed-horizon refinement; raw mean slopes span 0.041–0.389 (Table 4). On six scalar channels, we apply Chronos-Bolt-tiny's frozen frontend (Amazon, 2024) to identical 16-sample patches, with per-window normalization, at strides 1 (dense) and 16 (native). Native complete-patch outputs are a subset of dense outputs; this tests neither forecasting nor causal processing. All 29 dense windows have positive slopes, largely lost after shuffling (Figure 1, Table 1). Weather and oil temperature retain positive native-stride slopes; appliances and ETT load are weaker. Raw delay patches often have larger slopes, so learning alone does not explain the geometry. Fixed Lipschitz maps preserve Hölder bounds at fixed physical patch offsets, but overlap, scale selection, jumps, and mixed nested-grid seminorm trends qualify these findings (Appendices H.4 and H.6). SD is descriptive; the ETT channels share one source.

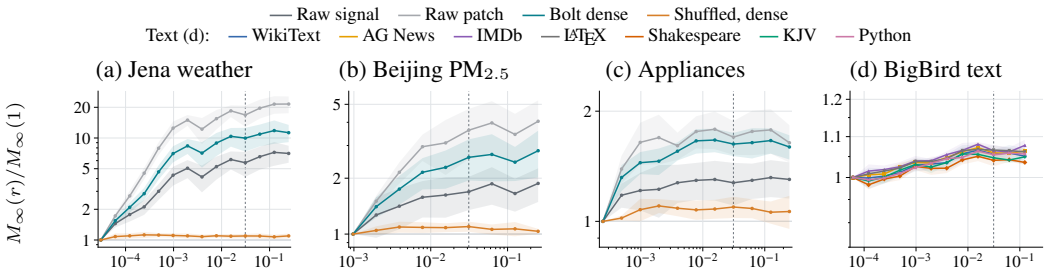

Figure 1: **Representation-dependent finite-scale regularity.** Worst-increment curves against normalized lag h (horizontal axis): physical time in (a)–(c), token position in (d). Curves are normalized per window by the finest-lag increment before averaging; shading is one sample SD. This preserves each window's slope and fixes the first point at one with zero SD. Raw-patch and shuffled controls probe patch geometry and temporal organization; LaTeX and Python are technical text controls. Dashed lines mark the fit cutoff; all panels have lower ordinate 0.85.

Text as a contrasting input model. Across seven text sources, BigBird input embeddings (Zaheer et al., 2020) with interpolated positions have nearly flat worst-increment curves at $N = 16{,}384$: mean slopes 0.011–0.013 (Appendix H.7). Appending words is not refinement of a fixed signal, and persistent order-one adjacent jumps would preclude a common Hölder bound. These finite-scale results do not support that hypothesis for this pipeline, but prove neither asymptotic failure nor impossibility for other representations. At native context lengths, BigBird's mean slope rises from 0.013–0.014 at input to 0.028–0.033 in intermediate blocks; this modest gain is non-monotone, also occurs after input shuffling, and does not persist to the final block. Appendix H.8 gives layer-wise curves and a causal DistilGPT2 comparison. Regularity therefore remains an assumption on the represented input family, not a conclusion of these fits.

Table 1: **Pretrained content comparison.** Fine-band $\widehat{\alpha}_\infty$, mean $\pm$ sample SD across m windows. Physical rows use Bolt content; WikiText-103 uses BigBird input. Shuffling precedes patching. Native dashes denote insufficient scales; the text shuffled control is unmeasured.

Domain	m	Input/dense	Native	Shuffled
Jena weather	8	.335±.037	.167±.040	.005±.003
Beijing $PM_{2.5}$	6	.257±.085	–	.023±.016
Appliances	4	.087±.010	.020±.017	.015±.012
Road traffic	3	.303±.009	–	.035±.011

Domain	m	Input/dense	Native	Shuffled
Oil temp.	4	.253±.055	.175±.041	.023±.004
ETT load	4	.181±.034	.043±.055	.026±.014
WikiText-103	4	.0122±.0012	n/a	–

4 Conclusion

Continuously extendable causal families are exactly those uniformly approximable across all resolutions by a length-independent masked transformer, whose parameters also approximate the continuous path limit. The proof represents finite prefixes and continuous histories by position-tagged laws on one compact space, then separates them with attention probes. The β-smooth target condition yields parameter-error rates and, on normalized target classes, finite-precision bounds. Bounded-weight empirical risk minimization gives logarithmic prediction rates from independent labeled sequences without a maximum-length factor; matching minimax exponents remain open. The positional obstruction confirms that masking alone cannot recover the normalized clock. Empirically, continuous pretrained features show dataset-dependent finite-scale regularity, whereas text input embeddings have nearly flat worst increments. Sampling and jump controls qualify these diagnostics, which do not certify a common all-resolution Hölder bound. Future work includes geometry-adaptive rates, efficient width–depth tradeoffs, and learning guarantees for dependent samples and computationally feasible optimization.

REPRODUCIBILITY STATEMENT

Section 2 states the assumptions and architectures; Appendices A–G give the complete mathematical arguments. Appendix H documents the empirical protocol, data windows, pretrained checkpoints, random seeds, scale choices, controls, and the limits of re-acquiring the seismic baseline. The anonymous numerical supplement provides saved measurements and figure-regeneration scripts; no model training is needed to regenerate the figures from these measurements.

AI USE STATEMENT

OpenAI Codex was used to develop and audit theoretical arguments, formulate and revise mathematical statements and proofs, design and refine empirical diagnostics, implement experimental and analysis code, interpret results, prepare figures, edit and restructure the manuscript, check literature and bibliography entries, and run compilation and visual diagnostics. The authors manually reviewed all AI-assisted work and take responsibility for the final mathematical arguments, empirical claims, text, code, and bibliography.

REFERENCES

Patrice Abry, Stéphane Jaffard, and Herwig Wendt. Irregularities and scaling in signal and image processing: Multifractal analysis. In Michael Frame and Nathan Cohen (eds.), *Benoit Mandelbrot: A Life in Many Dimensions*, pp. 31–116. World Scientific, 2015. doi: 10.1142/9789814366076_0003. URL https://arxiv.org/abs/1210.0482.

Beatrice Acciaio, Anastasis Kratsios, and Gudmund Pammer. Designing universal causal deep learning models: The geometric (hyper)transformer. *Mathematical Finance*, 34(2):671–735, 2024. doi: 10.1111/mafi.12389.

Amazon. Chronos-Bolt-Tiny: Pretrained time-series model and input-patch embedding. Hugging Face model card and checkpoint, 2024. URL https://huggingface.co/amazon/chronos-bolt-tiny. Checkpoint revision 93a8129; input embedding implementation in chronos-forecasting v2.0.0. Accessed 5 September 2026.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=gerNCVqqtR.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. URL https://arxiv.org/abs/2004.05150.

Adam Block, Zeyu Jia, Yury Polyanskiy, and Alexander Rakhlin. Intrinsic dimension estimation using Wasserstein distance. *Journal of Machine Learning Research*, 23(313):1–37, 2022. URL https://jmlr.org/papers/v23/21-1483.html.

Luis Candanedo. Appliances energy prediction. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5VC8G. Dataset.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840, 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. URL https://arxiv.org/abs/2306.15595.

Song Chen. Beijing PM2.5. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C5JS49. Dataset.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995. doi: 10.1109/72.392253.

Christa Cuchiero, Philipp Schmocker, and Josef Teichmann. Global universal approximation of functional input maps on weighted spaces. *Constructive Approximation*, 63(2):537–612, 2026. doi: 10.1007/s00365-025-09726-3.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019. doi: 10.18653/v1/P19-1285.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16344–16359, 2022. doi: 10.52202/068431-1189.

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 1993. doi: 10.1007/978-3-662-02888-9.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. LongRoPE: Extending LLM context window beyond 2 million tokens. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11091–11104, 2024. URL https://proceedings.mlr.press/v235/ding24i.html.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, 1993. doi: 10.1016/S0893-6080(05)80125-X.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c9028f7874df04843e7bf435ee4cd3c3-Abstract-Conference.html.

Luca Galimberti, Anastasis Kratsios, and Giulia Livieri. Designing universal causal deep learning models: The case of infinite-dimensional dynamical systems from stochastic analysis. *Constructive Approximation*, 2026. doi: 10.1007/s00365-026-09745-8. Online first.

Robert E. Gaunt and Siqi Li. Bounding Kolmogorov distances through Wasserstein and related integral probability metrics. *Journal of Mathematical Analysis and Applications*, 522(1):126985, 2023. doi: 10.1016/j.jmaa.2022.126985.

Tilmann Gneiting, Hana Ševčíková, and Donald B. Percival. Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statistical Science*, 27(2):247–277, 2012. doi: 10.1214/11-STS370. URL https://arxiv.org/abs/1101.1444.

Lukas Gonon and Juan-Pablo Ortega. Fading memory echo state networks are universal. *Neural Networks*, 138:10–13, 2021. doi: 10.1016/j.neunet.2021.01.025.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *The Annals of Applied Probability*, 33(1):28–69, 2023. doi: 10.1214/22-AAP1806.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018. doi: 10.1016/j.neunet.2018.08.025.

Joshua Hanson and Maxim Raginsky. Universal approximation of input–output maps by temporal convolutional nets. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14071–14081, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/39555391eb0624a439c5131b1bb8a2e0-Abstract.html.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

John Hogue. Metro interstate traffic volume. UCI Machine Learning Repository, 2019. URL https://doi.org/10.24432/C5X60B. Dataset.

Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/928170bcb050fe64a63fe781b82265aa-Abstract-Conference.html.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68926–68955, 2024. doi: 10.52202/079017-2202.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/96ee35c170cb8720033b16259c305da9-Abstract-Conference.html.

Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *Advances in Neural Information Processing Systems*, volume 37, pp. 115652–115681, 2024. doi: 10.52202/079017-3673.

Jason M. Klusowski and Andrew R. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory*, 64(12): 7649–7656, 2018. doi: 10.1109/TIT.2018.2874447.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Nikola B. Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data complexity estimates for operator learning, 2024. URL https://arxiv.org/abs/2405.15992. arXiv:2405.15992, version 2.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research*, 25(24): 1–67, 2024. URL https://jmlr.org/papers/v25/22-0719.html.

Shengjie Luo, Shanda Li, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. Your transformer may not be as powerful as you expect. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4301–4315, 2022. doi: 10.52202/068431-0311.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Max Planck Institute for Biogeochemistry. Jena Climate Dataset: Weather-station observations, 2009–2016. https://storage.googleapis.com/tensorflow/tf-keras-datasets/jena_climate_2009_2016.csv.zip, n.d. Accessed 24 August 2026.

Alexander Meister. Optimal classification and nonparametric regression for functional data. *Bernoulli*, 22(3):1729–1744, 2016. doi: 10.3150/15-BEJ709. URL https://arxiv.org/abs/1603.09130.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Erick Andres Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, 2020. doi: 10.1088/1361-6579/abc960.

Erick Andres Perez Alday, Annie Gu, Amit Shah, Chengyu Liu, Ashish Sharma, Salman Seyedi, Ali Bahrami Rad, Matthew Reyna, and Gari Clifford. Classification of 12-lead ECGs: The PhysioNet/Computing in Cardiology Challenge 2020. *PhysioNet*, July 2022. doi: 10.13026/dvyd-kd57. Version 1.0.2.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=R8sQPpGCv0.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Michaël E. Sander and Gabriel Peyré. Towards understanding the universality of transformers for next-token prediction. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d846c59be138a704e800f36e7fcb696a-Abstract-Conference.html.

Chang hoon Song, Geonho Hwang, Jun ho Lee, and Myungjoo Kang. Minimal width for universal property of deep RNN. *Journal of Machine Learning Research*, 24(121):1–41, 2023. URL https://jmlr.org/papers/v24/22-1191.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33416–33447, 2023. URL https://proceedings.mlr.press/v202/takakura23a.html.

Vilmos Totik. Polynomial approximation in several variables. *Journal of Approximation Theory*, 252:105364, 2020. doi: 10.1016/j.jat.2019.105364.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

James Vuckovic, Aristide Baratin, and Rémi Tachet des Combes. A mathematical theory of attention. *arXiv preprint arXiv:2007.02876*, 2020. URL https://arxiv.org/abs/2007.02876.

Haotong Yang, Fanxu Meng, Zhouchen Lin, and Muhan Zhang. Parrot Mind: Towards explaining the complex task reasoning of pretrained large language models with template-content structure. *arXiv preprint arXiv:2310.05452*, 2023. URL https://arxiv.org/abs/2310.05452. Revised 5 April 2024.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=ByxRM0Ntvr.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. O(n) connections are expressive enough: Universal approximability of sparse transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13783–13794, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/9ed27554c893b5bad850a422c3538c15-Abstract.html.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c8512d142a2d849725f31a9a7a361ab9-Abstract.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

A POSITIONAL ENCODING IS NEEDED FOR UNIVERSALITY

The theorems place normalized position in the hidden state but use no explicit length channel. We show that total length and score-side position cannot replace a position-revealing operation outside the normalized score channel. The obstruction is independent of the admissible modulus: constant sequences and paths belong to every class considered above.

Definition A.1 (Score-only positional normalized architecture)**.** At a fixed length n, a score-only positional normalized transformer has finite hidden dimensions and finite depth, and initializes according to

$$u_i^0 := \chi_n(z_i), \tag{25}$$

where the same map χ_n is used at every position. The map may depend on the known total length n, but not on i. Each normalized attention sublayer has the form

$$\mathcal{A}_{\ell,n}(u)_i := u_i + \sum_{r=1}^{H_\ell} W_{\ell,r} \sum_{j=1}^{i} \omega_{ij}^{\ell,r}(u_{1:i}) V_{\ell,r} u_j, \qquad \omega_{ij}^{\ell,r} \in \mathbb{R}, \quad \sum_{j=1}^{i} \omega_{ij}^{\ell,r} = 1. \tag{26}$$

The weights may depend arbitrarily on n, i, j, on the causal hidden prefix, and on absolute or relative positions. Thus they include ordinary masked softmax, score biases such as ALiBi, and position-dependent query/key maps such as RoPE. The value and output matrices are shared across positions, every other sublayer is a shared pointwise map, and position is absent from the initialization, values, residual features, and pointwise maps. There is no distinguished start token or other position-dependent value input.

A score-only positional normalized temporal transformer initializes $u^0(t) := \chi_\infty(x(t))$ and replaces (26) by

$$\mathcal{A}_{\ell,\infty}(u)(t) := u(t) + \sum_{r=1}^{H_\ell} W_{\ell,r} \int_{[0,t]} V_{\ell,r} u(s)\, \mathrm{d}\rho_{\ell,r,t,u}(s), \tag{27}$$

where $\rho_{\ell,r,t,u}$ is any finite signed Borel measure on $[0,t]$ with $\rho_{\ell,r,t,u}([0,t]) = 1$, possibly depending on t, absolute or relative time, and the causal hidden path $u|_{[0,t]}$. Thus the integral is well-defined against the current bounded Borel value path. Admissible kernels are required only to make every displayed integral well-defined and every sublayer map bounded Borel paths to bounded Borel paths. All non-attention maps remain pointwise and independent of t. No continuity in the parameter t is imposed on the resulting output: the lower bound below holds for this broader Borel-path class, and therefore also for any subclass required to map continuous paths to continuous paths.

Even upon allowing the architecture and parameters to vary with resolution, this class cannot recover the normalized clock on constant inputs.

Theorem A.2 (Sharp clock obstruction when position is confined to scores)**.** *Fix an admissible common modulus ω, with input classes X_n^ω and X_∞^ω as in (11) and (12), and consider the continuously extendable causal clock family*

$$\Theta_n(z)_i := i/n, \qquad \Theta_\infty(x)(t) := t. \tag{28}$$

Then the following two statements hold.

(a) Let $(S_n)_{n\geq 1}$ be any family in which each S_n is a scalar-output score-only positional normalized transformer at length n in the sense of Definition A.1. The architectures and parameters may vary arbitrarily with n. For every $c \in \Omega$ and $n \geq 1$, there is a scalar $b_n(c)$ such that

$$S_n(c, \ldots, c)_i = b_n(c), \qquad i \in [n], \tag{29}$$

and

$$\sup_{z \in X_n^\omega} \max_{i \in [n]} |S_n(z)_i - i/n| \geq \frac{1 - 1/n}{2}, \qquad n \geq 1, \tag{30}$$

and consequently

$$\sup_{n \geq 1} \sup_{z \in X_n^\omega} \max_{i \in [n]} |S_n(z)_i - i/n| \geq \frac{1}{2}. \tag{31}$$

(b) Every scalar-output score-only positional normalized temporal transformer S_∞ satisfies

$$\sup_{x \in X_\infty^\omega} \sup_{t \in [0,1]} |S_\infty(x)(t) - t| \geq \frac{1}{2}. \tag{32}$$

The per-resolution bound (30) and both all-resolution lower bounds are sharp.

The obstruction is constant-input invariance: every unit-mass normalized attention aggregate, even with signed and position-dependent weights, maps identical values to an identical value, and shared pointwise layers preserve this property. Comparing the first and last clock values yields the fixed-resolution bound and its sharp all-resolution limit. The claim is architecture-specific: distinguished start tokens, position-dependent values or residual features, position-dependent pointwise maps, unnormalized prefix sums, and distribution-specific guarantees lie outside its scope. Together with Theorems 2.4 and F.1, it shows that normalized residual position is sufficient, whereas masking, total length, and score-side position alone are not.

A.1 Proof

The proof uses a blockwise invariance of normalized affine aggregation. It holds in arbitrary hidden dimension, with any finite number of heads and any real, position-dependent unit-mass weights.

Lemma A.3 (Constant-state invariance)**.** *Consider a finite sequence whose hidden vectors are all equal: $u_1 = \cdots = u_n = v$. Then every output of a score-only positional normalized attention block (26) is the same vector,*

$$\mathcal{A}_{\ell,n}(u)_i = v + \sum_{r=1}^{H_\ell} W_{\ell,r} V_{\ell,r} v, \qquad i \in [n]. \tag{33}$$

A shared pointwise map also preserves equality across positions. The temporal counterparts preserve paths that are constant in time.

Proof. Finite sequences. Fix a head r and a query position i. Every value is $V_{\ell,r}v$, while the normalized weights may be completely different across keys and query positions. Nevertheless,

$$\sum_{j=1}^{i} \omega_{ij}^{\ell,r}(u_{1:i}) V_{\ell,r} v = \left(\sum_{j=1}^{i} \omega_{ij}^{\ell,r}(u_{1:i}) \right) V_{\ell,r} v = V_{\ell,r} v.$$

The result is independent of the prefix size and of every score-side positional mechanism. Summing the heads and adding the position-independent residual gives (33). Applying the same pointwise function to identical vectors again gives identical vectors.

Continuous time. For a constant path $u(t) = v$, each $\rho_{\ell,r,t,u}$ in (27) has total mass one, so

$$\int_{[0,t]} V_{\ell,r} v \, \mathrm{d}\rho_{\ell,r,t,u} = V_{\ell,r} v$$

also at $t = 0$. Thus the temporal block produces the same constant path. Induction proves the claim through any finite composition. □

Proof of Theorem A.2. Discrete lower bound. For a constant input $(c, \ldots, c)$, the shared initialization χ_n produces a constant hidden sequence. Lemma A.3 and induction through the length-n architecture prove (29), even when every attention sublayer uses position-dependent scores. This argument is separate at each n, so neither the parameters nor the architecture need be shared across lengths.

For every scalar b, the triangle inequality at the first and last positions gives

$$|1 - 1/n| \leq |b - 1/n| + |b - 1| \leq 2 \max_{i \in [n]} |b - i/n| . \tag{34}$$

Using $b = b_n(c)$, then taking the supremum over $z \in X_n^\omega$, proves (30). Taking the supremum over n proves (31).

Continuous lower bound and sharpness. In continuous time, Lemma A.3 gives a constant output b on every constant input path, and $1 \le |b| + |b-1|$ proves (32).

The clock family is continuously extendable: all finite maps are continuous and prefix-causal, the boundary map is continuous, and its joint evaluation is $\overline{\Theta}(\xi) = \tau(\xi)$, which is continuous by (150). At each fixed length n, the constant predictor $S_n \equiv (1 + 1/n)/2$ attains an error exactly equal to $(1 - 1/n)/2$, proving sharpness of (30). The constant shared predictor $S \equiv 1/2$, implemented by zero attention contributions and a final shared affine bias, attains all-resolution error $1/2$ in both discrete and continuous time. Hence both all-resolution bounds are sharp as well. □

B QUANTITATIVE APPROXIMATION RATES

The upper bound holds on every compact all-resolution class generated by an admissible common modulus. The rate comes instead from regularity of the target on causal histories. We therefore fix an arbitrary admissible modulus ω and write

$$X_n := X_n^{\omega}, \qquad X_\infty := X_\infty^{\omega}. \tag{35}$$

Let $\mathfrak{X}_\omega$ be the corresponding completed state space from (149)–(150), and let $\mathcal{M}_\omega$ be its compact prefix-law image from Proposition D.5. These constructions are summarized in the detailed proof guide in Appendix D.1 and proved in Appendix D. In the upper-bound statement and proof, $\mathfrak{X}$ and $\mathcal{M}$ abbreviate these two spaces. We first prove one estimate at every discrete resolution and then transfer it, without loss, to the continuous case. Only after proving this generic-modulus result do we specialize to Hölder histories to investigate optimality.

B.1 β-SMOOTH TARGET REGULARITY

To turn qualitative density into a rate, we measure how much the target can vary between histories that finite collections of attention probes cannot yet distinguish. A single family of smooth-test distances covers both Hölder and higher-order regularity on the completed history space used in the qualitative proof. This section expands Definition 2.7; its normalized teacher class is the ball $R_{\beta;\omega} \le 1, \|f^\star\|_\infty \le 1$.

Causal-history factor. Recall from (167) that every continuously extendable causal family factors uniquely as

$$\overline{F}^\star(\xi) = f^\star(\mu_\xi), \qquad f^\star : \mathcal{M} \to \mathbb{R}^{d'}. \tag{36}$$

The map $f^\star$ is therefore the common readout of the discrete and continuous targets after histories with the same causal information have been identified.

Causal-history metric. Recall the continuous endpoint extractor from (165). Concretely,

$$\mathfrak{e}(\mu_\xi) := y_\xi := \begin{cases} (z_i, i/n), & \xi = (n, z, i), \\ (x(t), t), & \xi = (\infty, x, t). \end{cases} \tag{37}$$

Set $E := \Omega \times [0,1]$. Consequently

$$\iota : \mathcal{M} \to \mathcal{P}(E) \times E, \qquad \iota(\mu) := (\mu, \mathfrak{e}(\mu)), \tag{38}$$

is a homeomorphism onto its image. The augmented state is precisely the graph of $\mathfrak{e}$, hence is canonically identified with $\mathcal{M}$.

Equip E with

$$d_E((x,s),(y,t)) := \|x - y\|_\infty + |s - t|, \tag{39}$$

and let W_1 be the associated 1-Wasserstein distance. Set

$$D_{\mathrm{caus}}(\mu,\nu) := W_1(\mu,\nu) + \|\mathfrak{e}(\mu) - \mathfrak{e}(\nu)\|_\infty, \qquad \mu,\nu \in \mathcal{M}. \tag{40}$$

Although topologically redundant, the endpoint term is essential for the rate: a positive-time continuous prefix assigns zero mass to its terminal token, whereas the residual coordinates retain that token exactly. Here the endpoint is the current token and its position, not necessarily the end of the full sequence. In our attention construction, the heads write only into auxiliary coordinates, so the residual connection carries these endpoint coordinates unchanged to the pointwise readout (Lemma D.10).

Isotropic Hölder restriction norm. Let $\mathsf{Q} := [-1,1]^d \times [0,1]$, and extend d_E to Q by the same formula as in (39). For $\sigma > 0$, put $r_\sigma := \lceil \sigma \rceil - 1$ and $\theta_\sigma := \sigma - r_\sigma \in (0,1]$. For a multi-index $a = (a_1, \dots, a_{d+1}) \in \mathbb{N}_0^{d+1}$, write $|a| := \sum_{j=1}^{d+1} a_j$ and $\partial^a := \partial_1^{a_1} \cdots \partial_{d+1}^{a_{d+1}}$. We use $C^\sigma(\mathsf{Q}) := C^{r_\sigma,\theta_\sigma}(\mathsf{Q})$: its elements are real-valued functions whose derivatives through order r_σ exist in the interior and extend continuously to the boundary, with finite full isotropic norm

$$\begin{aligned} \|g\|_{C^\sigma(\mathsf{Q})} &:= \sum_{|a| \le r_\sigma} \|\partial^a g\|_{L^\infty(\mathsf{Q})} + \sum_{|a| = r_\sigma} [\partial^a g]_{\theta_\sigma;\mathsf{Q}}, \\ [h]_{\theta;\mathsf{Q}} &:= \sup_{\substack{u,v \in \mathsf{Q} \\ u \ne v}} \frac{|h(u) - h(v)|}{d_E(u,v)^\theta}. \end{aligned} \tag{41}$$

Here $\partial^0 g = g$, so this also defines the norm when $r_\sigma = 0$. In particular, at integer orders we use the endpoint convention $C^m(\mathsf{Q}) = C^{m-1,1}(\mathsf{Q})$, rather than the classical space defined only by continuous derivatives through order m. For a function $\varphi : E \to \mathbb{R}$, the restriction norm is

$$\|\varphi\|_{C^\sigma(E)} := \inf\left\{\|g\|_{C^\sigma(\mathsf{Q})} : g \in C^\sigma(\mathsf{Q}), \quad g|_E = \varphi\right\}, \tag{42}$$

with infimum $+\infty$ if there is no such extension; $C^\sigma(E)$ is the space where this norm is finite. Equivalent norms on this fixed finite-dimensional box change the estimates only by constants depending on dimension and the fixed smoothness order.

Smooth-test Wasserstein distances. Testing against increasingly smooth functions gives a metric scale adapted to the moments recovered by attention. This is the Hölder integral probability metric construction (Block et al., 2022); at integer orders it is equivalent, up to norm conventions, to bounded smooth Wasserstein metrics (Gaunt & Li, 2023). Define

$$d_\sigma(\mu, \nu) := \sup_{\|\varphi\|_{C^\sigma(E)} \le 1} \int_E \varphi \, \mathrm{d}(\mu - \nu). \tag{43}$$

The supremum is unchanged if the integral is replaced by its absolute value, since the test class is closed under $\varphi \mapsto -\varphi$. Because restrictions of polynomials belong to $C^\sigma(E)$ and are dense in $C(E)$, this test class separates probability measures and d_σ is a metric on $\mathcal{P}(E)$.

For $\beta > 0$, put $\beta_- := \min\{\beta, 1\}$ and $\beta_+ := \max\{\beta, 1\}$. Define

$$W_{1,\beta}(\mu, \nu) := d_{\beta_+}(\mu, \nu)^{\beta_-}, \qquad \delta_e(\mu, \nu) := \|\mathfrak{e}(\mu) - \mathfrak{e}(\nu)\|_\infty, \tag{44}$$

and augment the law distance by the difference of the retained endpoint coordinates:

$$\Delta_\beta(\mu, \nu) := W_{1,\beta}(\mu, \nu) + \delta_e(\mu, \nu)^{\beta_-}. \tag{45}$$

In $W_{1,\beta}$, β is a smoothness index: it denotes neither a power of W_1 nor a Wasserstein transport order. The following lemma shows that these distances retain the original compact topology while changing its quantitative geometry.

Lemma B.1 (Metric properties and Wasserstein normalization)**.** *For every $\beta > 0$, $W_{1,\beta}$ is a metric on $\mathcal{P}(E)$ inducing weak convergence, and Δ_β is a metric on $\mathcal{M}$ inducing its original compact topology. With constants depending only on (d, β),*

$$W_{1,1} = d_1 \asymp W_1, \qquad \Delta_\beta \asymp D^\beta_{\mathrm{caus}} \quad (0 < \beta \le 1), \qquad \Delta_\beta = d_\beta + \delta_e \quad (\beta > 1). \tag{46}$$

If the test norm at order one is replaced by the Lipschitz seminorm for d_E, then $W_{1,1} = W_1$ exactly.

Proof. For $\sigma \ge 1$, the full C^σ norm controls the Lipschitz seminorm, so Kantorovich duality gives $d_\sigma \le C_{d,\sigma} W_1$. Conversely, a one-Lipschitz test on E extends to Q with the same Lipschitz constant. Subtracting its value at a fixed point bounds its uniform norm by $\mathrm{diam}_{d_E}(\mathsf{Q})$, without changing its integral against $\mu - \nu$. Its full C^1 restriction norm is therefore bounded by a dimension-dependent constant, proving $W_1 \le C_d d_1$. With the Lipschitz seminorm alone, Kantorovich duality gives exact equality.

For $\sigma \ge 1$, d_σ is a separating dual metric, and $d_\sigma \lesssim W_1$ makes the identity from the weakly compact space $\mathcal{P}(E)$ to its d_σ-metric topology continuous. Compactness and the Hausdorff property imply that these topologies coincide. For $q \in (0, 1]$, the inequality $(a + b)^q \le a^q + b^q$ shows that the q-th power of a metric is again a metric with the same topology. Taking $q = \beta_-$ proves the assertion for $W_{1,\beta}$. The endpoint extractor is continuous on $\mathcal{M}$, so $\delta_e^{\beta_-}$ is a continuous pseudometric; adding it preserves the metric and topology. Finally, for $0 < \beta \le 1$,

$$(a + b)^\beta \le a^\beta + b^\beta \le 2^{1-\beta}(a + b)^\beta,$$

which, together with $d_1 \asymp W_1$, proves the comparison with D^β_{caus}. For $\beta > 1$, the equality is immediate. □

Target regularity. We measure regularity of the target by Lipschitz continuity in Δ_β. The preceding comparison recovers ordinary Hölder continuity at low orders. At high orders, smooth test functions replace powers of the underlying distance: using $D_{\mathrm{caus}}^{\beta}$ for $\beta > 1$ would generally violate the triangle inequality and force a target satisfying that modulus to be constant along every rectifiable path. Increasing $\beta > 1$ instead shrinks the smooth-test metric, up to fixed norm-embedding constants, and hence strengthens target regularity. The endpoint remains first order because the residual coordinates preserve it exactly.

Definition B.2 (β-smooth regularity on causal histories)**.** Fix $\beta > 0$. The β-smooth seminorm of a completed causal family $\mathbf{F}^\star$ on the class determined by ω is

$$R_{\beta;\omega}(\mathbf{F}^\star) := \sup_{\substack{\mu,\nu\in\mathcal{M}_\omega \\ \mu\neq\nu}} \frac{\|f^\star(\mu) - f^\star(\nu)\|}{\Delta_\beta(\mu,\nu)}. \tag{47}$$

Equivalently, $R_{\beta;\omega}(\mathbf{F}^\star)$ is the optimal (smallest) Lipschitz constant of $f^\star : (\mathcal{M}_\omega, \Delta_\beta) \to \mathbb{R}^{d'}$, with value $+\infty$ if no finite Lipschitz constant exists. The family is *β-smooth* when this quantity is finite.

For $0 < \beta \leq 1$, this is equivalent to ordinary β-Hölder continuity in the causal-history metric. For noninteger $\beta = m + \vartheta > 1$, with $m \in \mathbb{N}$ and $0 < \vartheta < 1$, the measure term tests isotropic $C^{m,\vartheta}$ regularity in the joint token–time variable. At an integer $\beta = m \geq 2$, we use the endpoint convention $C^{m-1,1}$. In both high-order regimes, the endpoint contribution to Δ_β is the first-order term δ_e, because these coordinates are retained exactly rather than reconstructed from attention statistics. This does not require higher differentiability of endpoint-only functions. Thus β-smoothness means stability in the smooth-test metric, not generic Fréchet C^β regularity of a path operator.

The order β is fixed while the parameter budget grows: constants may depend nonuniformly on β, and we make no analytic-class claim. Moreover, because $\mathcal{M} = \mathcal{M}_\omega$ depends on the input class, both $R_{\beta;\omega}(\mathbf{F}^\star)$ and the learned readout may depend on ω, even though the approximation constants below do not.

B.2 Quantitative all-resolution rate

We now strengthen Theorem 2.4 by converting target regularity into an explicit parameter–error tradeoff, uniform over all sequence lengths. For a completed family $\mathbf{F}^\star = ((F_n^\star)_{n\geq 1}, F_\infty^\star)$ and a shared transformer with its temporal realization, define the discrete and boundary errors by

$$E_{\mathrm{disc}}(T;\mathbf{F}^\star) := \sup_{n\geq 1}\sup_{z\in X_n}\max_{i\in[n]} \|T(\phi_n(z))_i - F_n^\star(z)_i\|, \tag{48}$$

$$E_\infty(T;\mathbf{F}^\star) := \sup_{x\in X_\infty} \|T_\infty(\phi_\infty(x)) - F_\infty^\star(x)\|_\infty. \tag{49}$$

Their maximum is the joint all-resolution error:

$$\mathrm{Err}(T;\mathbf{F}^\star) := \max\{E_{\mathrm{disc}}(T;\mathbf{F}^\star), E_\infty(T;\mathbf{F}^\star)\}. \tag{50}$$

For a transformer with H scalar heads, residual width $r = d + H + 2$, and an M-unit pointwise readout, the *dense parameter count* counts every matrix and bias entry, including structural zeros, rather than only nonzero coefficients:

$$\mathrm{par}(T) := 4Hr + (r + d' + 1)M + d'. \tag{51}$$

Each head contributes four matrices with r entries; the two readout matrices and two biases contribute $rM + M + d'M + d'$. Thus structural zeros and biases are charged, shared entries are charged once, and the fixed lift $\mathcal{E}_H$ is parameter-free. This count charges neither parameter magnitude nor numerical precision. The following theorem records the joint discrete–continuous rate, the head budget, and the equivalent accuracy-to-parameter bound. Its joint conclusion combines the discrete construction with the lossless boundary transfer proved in Corollary B.4.

Theorem B.3 (Quantitative all-resolution causal universality)**.** *Fix an admissible common modulus ω and $\beta > 0$, and let X_n and X_∞ be the input classes in (35). Let $(F_n^\star)_{n\geq 1}$ be a continuously extendable causal family in the sense of Definition 2.2, with its unique extension $F_\infty^\star$, and write $\mathbf{F}^\star = ((F_n^\star)_{n\geq 1}, F_\infty^\star)$. Assume $\mathbf{F}^\star$ is β-smooth in the sense of Definition B.2, so that $R_{\beta;\omega}(\mathbf{F}^\star) < \infty$. There exist a constant $C_\beta = C(d,d',\beta) > 0$ and an integer $p_0 = p_0(d,d',\beta) \geq 16$, both independent*

of ω and the target. For every integer $p \geq p_0$, there are integers $H_p, M_p \geq 1$ and a transformer T_p of the shallow form (7)–(9), with residual width $d + H_p + 2$, H_p scalar heads, and an M_p-unit pointwise readout. Its dimensions, matrices, and biases are independent of n, and, together with its temporal realization, it satisfies

$$\mathrm{par}(T_p) \leq p, \qquad \mathrm{Err}(T_p; \mathbf{F}^\star) \leq C_\beta R_{\beta;\omega}(\mathbf{F}^\star) \left(\frac{\log\log p}{\log p} \right)^{\beta/(d+2)}. \tag{52}$$

The attention block can be chosen with

$$H_p \leq C_\beta \left(\frac{\log p}{\log\log p} \right)^{(d+1)/(d+2)} \tag{53}$$

scalar heads, embedding width $d + 2 + H_p$, and every attention-matrix entry in the interval $[-1, 1]$. Equivalently, if $R_{\beta;\omega}(\mathbf{F}^\star) > 0$, then for every $0 < \varepsilon \leq R_{\beta;\omega}(\mathbf{F}^\star)$ the construction attains joint discrete–continuous error at most ε using

$$\begin{aligned} \log p_\varepsilon &\leq C_\beta \left(\frac{R_{\beta;\omega}(\mathbf{F}^\star)}{\varepsilon} \right)^{(d+2)/\beta} \log\left(2 + \frac{R_{\beta;\omega}(\mathbf{F}^\star)}{\varepsilon} \right), \\ H_\varepsilon &\leq C_\beta \left(\frac{R_{\beta;\omega}(\mathbf{F}^\star)}{\varepsilon} \right)^{(d+1)/\beta}. \end{aligned} \tag{54}$$

If $R_{\beta;\omega}(\mathbf{F}^\star) = 0$, the target factor is constant and a final affine bias realizes it exactly.

The uniformity is over sequence lengths, not over input classes: the learned transformer and $R_{\beta;\omega}(\mathbf{F}^\star)$ may depend on ω, whereas the exponent, constants, threshold p_0, and attention-entry bound do not. The parameter count includes the full one-hidden-layer readout, whose width M_p generally grows much faster than the head count H_p. The denominator $d + 2 = (d + 1) + 1$ combines $O(K^{d+1})$ joint token–time probes with an $O(K \log K)$ stable-recovery cost in the logarithmic readout complexity. Target regularity first reduces the oscillation on each feature cell to $O(K^{-\beta})$; a merely Lipschitz shallow readout is fitted only after this reduction, so it does not saturate the rate at $\beta = 1$. The theorem counts real parameters; Corollary B.5 below also controls magnitudes and finite precision on a normalized target ball.

The following transfer principle shows why the discrete construction already controls the continuous boundary. Its error identity does not require target regularity or the quantitative theorem.

Corollary B.4 (Quantitative rate on the path boundary)**.** *Fix an admissible common modulus ω, let $(F_n^\star)_{n\geq 1}$ satisfy Definition 2.2, and write $\mathbf{F}^\star = ((F_n^\star)_{n\geq 1}, F_\infty^\star)$ for its completed family. Every shared causal transformer T covered by Proposition F.2, together with its temporal realization, satisfies*

$$E_\infty(T; \mathbf{F}^\star) \leq E_{\mathrm{disc}}(T; \mathbf{F}^\star), \qquad \mathrm{Err}(T; \mathbf{F}^\star) = E_{\mathrm{disc}}(T; \mathbf{F}^\star). \tag{55}$$

Consequently, for a β-smooth family, the transformer T_p constructed for Theorem B.3, equipped with the temporal attention defined by the same parameters, obeys

$$\mathrm{Err}(T_p; \mathbf{F}^\star) \leq C_\beta R_{\beta;\omega}(\mathbf{F}^\star) \left(\frac{\log\log p}{\log p} \right)^{\beta/(d+2)}. \tag{56}$$

The bound $\mathrm{par}(T_p) \leq p$, the head estimate (53), and the accuracy form (54) are unchanged for this simultaneous discrete–continuous estimate.

Proof. Fix $x \in X_\infty$, put $z^n := \mathcal{S}_n x$, and denote $e := E_{\mathrm{disc}}(T; \mathbf{F}^\star)$. At every finite resolution,

$$\| \mathcal{I}_n(T(\phi_n(z^n)) - F_n^\star(z^n)) \|_\infty \leq e,$$

because affine interpolation is a convex combination of grid errors. Proposition F.2 sends the first interpolated term to $T_\infty(\phi_\infty(x))$. By compatibility, the distance from the second term to $F_\infty^\star(\mathcal{I}_n \mathcal{S}_n x)$ tends to zero. Since $\mathcal{I}_n \mathcal{S}_n x \to x$ and $F_\infty^\star$ is continuous, the second term therefore converges to $F_\infty^\star(x)$. Passing to the limit and then taking the supremum over x proves $E_\infty \leq E_{\mathrm{disc}}$, hence (55); no additional heads, parameters, or approximation loss are introduced. Applying this identity to the discrete construction in Appendix C yields (56) and the joint conclusion of Theorem B.3. □

On a normalized target ball, the same construction also admits a finite binary description. The next result charges precision separately from the dense parameter count.

Corollary B.5 (Finite precision on a normalized target ball)**.** *Under the assumptions of Theorem B.3, suppose also that $R_{\beta;\omega}(\mathbf{F}^\star) \le 1$ and $\|f^\star\|_\infty \le 1$. There are constants C_β, p_0, depending only on (d, d', β), such that, for every integer $p \ge p_0$, a shallow shared transformer $\widetilde{T}_p$ satisfies*

$$\mathrm{par}(\widetilde{T}_p) \le p, \qquad \mathrm{Err}(\widetilde{T}_p; \mathbf{F}^\star) \le C_\beta \left(\frac{\log\log p}{\log p} \right)^{\beta/(d+2)}. \tag{57}$$

Its attention entries lie in $[-1, 1]$, all other parameters have magnitude at most $p + 1$, and every parameter is dyadic and has a binary description of at most $C_\beta \log_2 p$ bits. The architecture can be fixed across the normalized target ball at each budget, and the total description length, including its dimensions, is at most $C_\beta p \log_2 p$.

The proof, given in Appendix C, tracks the ridge coefficients and rounds them on a common dyadic grid. This is a normalized absolute-error guarantee, not the homogeneous bound proportional to $R_{\beta;\omega}$: an independently sized additive offset may require additional bits when the seminorm is small, and an arbitrary constant target need not have an exact finite-bit representation. Only parameter storage is quantized; network evaluation is still analyzed in exact arithmetic. The result controls an existence construction, not a claim of conditioning independent of the parameter budget. Its bounded weights are also the approximation input to the empirical-risk guarantee in Theorem 2.9; Appendix G supplies the additional statistical argument and sampling assumptions.

B.3 Proof mechanism: quantitative resolution by causal probes

The proof follows the chain: attention probes recover moments, moments resolve the history metric, target regularity controls the unresolved fibers, and a counted ReLU readout approximates the resulting finite-dimensional map. The detailed estimates are in Appendix C; here are the five steps.

1. **Resolve causal histories by finitely many probes.** Put $D = d + 1$. The direction grid and sampling scales in (72)–(73) define $\Phi_K(\mu, \mathfrak{e}(\mu))$. It retains the current lifted token and uses $H_K = 2K(K+1)^{D-1} = O(K^{d+1})$ normalized log-Laplace-derivative samples. High-order interpolation in Lemma C.1 recovers every projected moment through degree K, and tensor interpolation recovers the mixed moments.
2. **Convert moments into metric resolution.** Lemma C.2 turns moment control into control of smooth token–position test functions. Proposition C.3 then gives, for τ_K-close feature vectors on the causal-history graph,
$$d_{\beta_+}(\mu, \nu) \le C_\beta K^{-\beta_+}, \qquad \delta_e(\mu, \nu) \le \tau_K,$$
where τ_K decays faster than any fixed inverse power of K.
3. **Control target oscillation on each feature cell.** Since $\beta_+\beta_- = \beta$, Step 2 gives $\Delta_\beta(\mu, \nu) \le C_\beta K^{-\beta}$. Definition B.2 therefore bounds the oscillation of $f^\star$ on a feature cell by $C_\beta R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}$. This is the step that uses the full order β; the later readout need only be Lipschitz.
4. **Realize the finite-dimensional readout.** The clipped infimal envelope (114) extends the approximately fiber-constant target with Lipschitz constant $O(R_{\beta;\omega}(\mathbf{F}^\star)/\tau_K)$. Lemma C.5 mollifies this envelope and applies a dimension-explicit shallow-ReLU estimate, giving a one-hidden-layer readout of width W_K with $\log W_K = O_\beta\left(\underbrace{K^{d+1}}_{\text{joint token–time features}} \underbrace{K \log(K+1)}_{\text{stable recovery}} \right)$.
5. **Count parameters and invert the scale relation.** Lemma D.10 computes all H_K probes exactly in one masked block, simultaneously on every finite grid and on the temporal boundary. Counting the attention matrices and shallow readout gives $\log \mathrm{par}(T_K) = O_\beta(K^{d+2} \log(K+1))$; inverting this relation gives (52) and (54). Thus $d + 2 = (d+1) + 1$: the exponent comes from feature dimension and stability, not from the input modulus. Proposition F.2 then gives Corollary B.4 without changing the rate.

B.4 Interpreting the smoothness condition

The definition is tailored to the information recovered by attention, but it can be checked on familiar classes of functionals. We give endpoint-only examples, a first-variation certificate, and Lipschitz readouts of smooth history statistics.

Endpoint-only maps. Any map $f^\star(\mu) \coloneqq g(\mathfrak{e}(\mu))$ with g Lipschitz satisfies the definition for every fixed $\beta > 1$. Membership at all such orders reflects the separate endpoint term, not higher classical smoothness of g. The generic construction can exploit a larger β by resolving the endpoint through increasingly fine feature cells, paid for by a very wide readout. Membership in this scale therefore does not assert classical differentiability of endpoint-only targets.

The endpoint is exactly determined by a continuous graph law, but its extraction need not be Wasserstein-Lipschitz. For a concrete example, fix $0 < \alpha \le 1$, $L > 0$, and $0 < c \le \min\{L, 1\}$. The paths

$$x_h(s) \coloneqq c(s - 1 + h)_+^\alpha e_1, \qquad 0 < h < 1,$$

and the zero path belong to the same α-Hölder ball, where e_1 denotes the first token coordinate vector. Their terminal endpoints differ by ch^α, whereas coupling their terminal prefix laws at equal times gives

$$W_1(\mu_{\infty,x_h,1}, \mu_{\infty,0,1}) \le \int_0^1 \|x_h(s)\|_\infty \, \mathrm{d}s = \frac{c}{\alpha + 1} h^{\alpha+1}. \tag{58}$$

Thus retaining the endpoint coordinates supplies quantitative stability that does not follow from the law's exact information alone.

A sufficient first-variation criterion. At high order, Lemma C.4 provides a standard sufficient criterion through first variations. It suffices that $f^\star$ be the trace of an ambient extension that is Lipschitz in the explicitly retained token–position endpoint and whose linear functional derivative is uniformly C^β, modulo constants, in its integration variable E. Values of this extension away from the causal-history graph are auxiliary: they certify smoothness but do not redefine the target. For example, consider the smooth cylindrical functional

$$\widetilde{f}(\mu, y) \coloneqq \Upsilon\left(y, \int_E \varphi_1 \, \mathrm{d}\mu, \dots, \int_E \varphi_q \, \mathrm{d}\mu\right) \tag{59}$$

with jointly token–position C^β test functions and a smooth finite-dimensional outer map. Its functional derivative is a bounded linear combination of the φ_j's, and its dependence on y is Lipschitz on the relevant compact set. The ambient formulation is convenient but stronger than necessary: the proof only uses the derivative identity and its uniform bound along mixture segments joining pairs of laws in $\mathcal{M}_\omega$, together with endpoint Lipschitz control at the corresponding states.

Remark B.6 (Lipschitz readouts of smooth history statistics). The smooth cylindrical example can be verified without differentiating its outer map. If $\varphi_1, \dots, \varphi_q \in C^\beta(E)$, $\beta > 1$, and Ψ is Lipschitz on the relevant compact product, then

$$f(\mu) \coloneqq \Psi\left(\mathfrak{e}(\mu), \int_E \varphi_1 \, \mathrm{d}\mu, \dots, \int_E \varphi_q \, \mathrm{d}\mu\right)$$

is β-smooth. Indeed, each integral difference is at most $\|\varphi_j\|_{C^\beta(E)} \, d_\beta(\mu, \nu)$, and the endpoint is controlled separately. Writing e_1' for the first output coordinate vector, for example, $f(\mu) = \left|\int_E x_1 \, \mathrm{d}\mu\right| e_1'$ belongs at every fixed high order even though its outer absolute value is not differentiable. This is stability in the smooth-test metric, not classical smoothness of the entire operator.

In sequence form, these examples read

$$F_n(z)_i = \Psi\left((z_i, i/n), \frac{1}{i}\sum_{j=1}^{i} \varphi_1(z_j, j/n), \dots, \frac{1}{i}\sum_{j=1}^{i} \varphi_q(z_j, j/n)\right).$$

The path realization replaces each normalized sum by $t^{-1}\int_0^t \varphi_j(x(s), s) \, \mathrm{d}s$, with value $\varphi_j(x(0), 0)$ at zero.

B.5 Logarithmic lower bounds on Hölder histories

The upper rate is logarithmic in the parameter budget. To test whether such slow decay reflects the complexity of long histories, we now specialize the input modulus to

$$\omega_{\alpha,L}(r) := Lr^{\alpha}, \qquad 0 < \alpha \le 1, \quad L > 0, \tag{60}$$

and write

$$X_n^{\alpha,L} := X_n^{\omega_{\alpha,L}}, \qquad X_\infty^{\alpha,L} := X_\infty^{\omega_{\alpha,L}}. \tag{61}$$

Let $\mathfrak{X}_{\alpha,L}$ and $\mathcal{M}_{\alpha,L}$ denote the corresponding completion and prefix-law image, and abbreviate $R_{\beta;\alpha,L} := R_{\beta;\omega_{\alpha,L}}$. For every fixed $\beta > 0$, we prove a logarithmic worst-case lower bound when approximants have a finite binary description. This does not establish the optimality of the upper exponent, or a lower bound for a count of unrestricted real parameters. Appendix C.6 contains the proofs.

Target class and description budget. We use exactly the target class of Theorem B.3, normalized to separate regularity from amplitude:

$$\mathfrak{F}^{(1)}_{\beta;\alpha,L} := \left\{ \mathbf{F}^\star : R_{\beta;\alpha,L}(\mathbf{F}^\star) \le 1, \quad \sup_{\mu\in\mathcal{M}_{\alpha,L}} \|f^\star(\mu)\| \le 1 \right\}. \tag{62}$$

Here every $\mathbf{F}^\star$ is a continuously extendable causal family and $f^\star$ its canonical history factor. A B-bit description selects at most 2^B decoded maps, so the smallest worst-case error over all such descriptions is

$$\mathcal{A}_B^{\mathrm{code}}(\beta;\alpha,L) := \inf_{\substack{\mathcal{G}\subset(\mathbb{R}^{d'})^{\mathcal{M}_{\alpha,L}} \\ |\mathcal{G}|\le 2^B}} \sup_{\mathbf{F}^\star\in\mathfrak{F}^{(1)}_{\beta;\alpha,L}} \inf_{g\in\mathcal{G}} \sup_{\mu\in\mathcal{M}_{\alpha,L}} \|g(\mu) - f^\star(\mu)\|. \tag{63}$$

The decoder is fixed across targets but otherwise unrestricted, so this lower bound applies in particular to finite-precision transformers. Only the description is charged; no training, sampling, or noise model is assumed. The error is uniform on the same completed history space as in the upper theorem.

Many distinguishable histories force slow approximation. The key is a multilevel construction of Hölder paths whose prefix laws are well separated by the target metric Δ_β. Write

$$\mathsf{P}_{\beta;\alpha,L}(r) := \max\{|S| : S \subset \mathcal{M}_{\alpha,L},\ \Delta_\beta(\mu,\nu) \ge r \text{ for distinct } \mu,\nu\in S\}. \tag{64}$$

This number is finite by compactness. The construction below yields enough histories to encode more admissible target patterns than a B-bit description can distinguish.

Proposition B.7 (Logarithmic lower bound on Hölder histories)**.** *Fix $d, d' \ge 1$, $0 < \alpha \le 1$, $L > 0$, and $\beta > 0$. There are $c, r_0 > 0$, depending only on (d,α,L,β), such that*

$$\log \mathsf{P}_{\beta;\alpha,L}(r) \ge cr^{-1/\beta}\left[1 + (1-\alpha)\log(1/r)\right] \qquad (0 < r \le r_0). \tag{65}$$

Consequently, for some $c > 0$, all sufficiently large integers B satisfy

$$\mathcal{A}_B^{\mathrm{code}}(\beta;\alpha,L) \ge c\left(\frac{1 + (1-\alpha)\log\log(B+e^e)}{\log(B+e^e)}\right)^{\beta}. \tag{66}$$

The proof places many token levels in disjoint time cells and separates the resulting paths by smooth tests. On an N-point history packing, independent signed target values extend with controlled β-smooth seminorm, producing 2^N distinguishable targets; a B-bit description cannot resolve them all when $N > B$. The lower rate is $(\log B)^{-\beta}$ for $\alpha = 1$ and $(\log\log B/\log B)^{\beta}$ for each fixed $\alpha < 1$. Thus logarithmically slow worst-case decay is unavoidable at finite description length. Neither the packing exponent nor its logarithmic dependence on α are claimed to be sharp.

Comparison with the constructive rate. The finite-precision guarantee in Corollary B.5 uses $O(p\log p)$ bits. Choosing $p \asymp_\beta B/\log B$ therefore gives, for all sufficiently large B,

$$\mathcal{A}_B^{\mathrm{code}}(\beta;\alpha,L) \le C_{d,d',\beta}\left(\frac{\log\log B}{\log B}\right)^{\beta/(d+2)}. \tag{67}$$

The lower and upper bounds are both logarithmic, but their exponents β and $\beta/(d+2)$ do not match. In particular, the construction's dimension penalty is not known to be necessary. The constants in the lower bound may depend on (α, L), whereas the normalized upper-bound constant does not. These are finite-description comparisons, not a matching minimax result for unrestricted real-parameter networks.

Why VC dimension cannot transfer the bound. Let $\mathfrak{T}_p^{\mathrm{sh}}$ denote the shallow shared transformers in (7)–(9) with at most p dense parameters, each with its canonical temporal realization. One might try to replace B by a capacity bound for p real parameters. This route fails: on the infinite-dimensional path domain, the threshold class generated by a fixed-size shallow architecture already has infinite VC dimension. More precisely, let

$$v_p(\alpha, L) \coloneqq \mathrm{VCdim}\left\{\mu \mapsto \mathbf{1}_{\{(\widehat{T}(\mu))_1 > 0\}} : T \in \mathfrak{T}_p^{\mathrm{sh}}\right\}, \tag{68}$$

where $\widehat{T} : \mathcal{M}_{\alpha,L} \to \mathbb{R}^{d'}$ is the causal-history factor of the completed transformer.

Proposition B.8 (Infinite threshold capacity at fixed size)**.** *For every $0 < \alpha \le 1$ and $L > 0$,*

$$v_p(\alpha, L) = \infty \qquad \textit{whenever} \qquad p \ge p_\star \coloneqq 5(d+3) + 1 + 2d'. \tag{69}$$

This remains true if every attention and readout weight is restricted to $[-1, 1]$.

Corollary B.9 (Infinite threshold capacity on finite sequences)**.** *Fix $0 < \alpha \le 1$, $L > 0$, and $p \ge p_\star$. For every $N \ge 1$, there exist a common length n and N distinct sequences in $X_n^{\alpha,L}$ that are shattered by the terminal-output threshold class*

$$z \in X_n^{\alpha,L} \longmapsto \mathbf{1}_{\{(T(\phi_n(z))_n)_1 > 0\}}, \qquad T \in \mathfrak{T}_p^{\mathrm{sh}},$$

even with every weight in $[-1, 1]$. In particular, this threshold class has infinite VC dimension on the disjoint union of all finite input lengths.

The common length may grow with N; the corollary does not assert infinite capacity at one fixed resolution. Its proof samples the finite collection of smooth paths and preserves their strictly signed moments on a sufficiently fine common grid.

The obstruction comes from selecting a real-valued Laplace parameter, not from growing the width. One temporal head can threshold the Laplace moment $\int_0^1 e^{\vartheta s} x(s)\, \mathrm{d}s$, and the one-parameter family $\{e^{\vartheta s} : 0 < \vartheta < 1\}$ shatters arbitrarily large finite sets of smooth paths chosen for that set. Hence no finite bound on $v_p(\alpha, L)$, polynomial or otherwise, can convert the packing result into a lower bound for unrestricted real parameters. Bounding the weights does not fix this zero-threshold obstruction; even fixed-margin fat-shattering remains infinite if the final output scale is unrestricted. By contrast, a class described by B-bit strings has at most 2^B members and VC dimension at most B, consistent with the finite-precision guarantee in Corollary B.5. A meaningful capacity comparison must therefore charge another resource, such as parameter precision, output scale together with a margin, or input discretization dimension. The finite-code lower bound remains valid, but optimality of the real-parameter upper rate remains open.

This zero-threshold obstruction does not preclude the regression guarantee of Theorem 2.9. With both attention and readout weights bounded, the real-valued predictor class admits finite uniform covers at every positive accuracy, independently of context length (Lemma G.2). The shattering margins can tend to zero, so threshold capacity and bounded squared-loss estimation measure different notions of complexity.

C Proofs for quantitative rates and history-complexity bounds

This appendix follows the five-step proof chain summarized in Subsection B.3. Exact causal attention probes first resolve finitely many moments; polynomial approximation then converts this information into a metric resolution of causal histories. The target regularity turns that resolution into a $K^{-\beta}$ oscillation bound, and a counted shallow ReLU readout converts the latter into the stated parameter rate. We close with the optimality analysis on the Hölder specialization.

Put $D \coloneqq d + 1$, write $y = (x, s) \in E = \Omega \times [0, 1]$, and equip E with

$$d_E\big((x, s), (x', s')\big) \coloneqq \|x - x'\|_\infty + |s - s'|\,. \tag{70}$$

We denote by W_1 the corresponding Wasserstein distance on $\mathcal{P}(E)$. Changing between this norm and the Euclidean norm used elsewhere only changes dimension-dependent constants; dependence on the fixed dimensions d, d' is suppressed in constant subscripts. In particular, all constants below are independent of the sequence length and of the prefix at which the output is read. Fix the arbitrary

admissible common modulus ω from Theorem B.3, and work on $\mathcal{M} := \mathcal{M}_\omega$. The modulus only selects this compact subset of the ambient history space: neither the probe estimates nor the constants below depend on ω. We use the same causal-history factor $f^\star : \mathcal{M} \to \mathbb{R}^{d'}$ as in (36). Its quantitative chart is $\iota(\mu) = (\mu, \mathfrak{e}(\mu))$, defined in (38); the second coordinate is already available in the residual coordinates. On the ambient product $\mathcal{P}(E) \times E$, write

$$D_{\mathrm{prod}}\big((\mu, y), (\nu, y')\big) := W_1(\mu, \nu) + \|y - y'\|_\infty . \tag{71}$$

Its restriction to the graph of $\mathfrak{e}$ is exactly D_{caus} from (40).

C.1 A FINITE FAMILY OF EXACT MASKED-ATTENTION PROBES

The first step constructs a finite feature map that is computed exactly by masked attention and whose small cells have small diameter in the metrics relevant to the target.

We use the box $\mathsf{Q} = [-1, 1]^d \times [0, 1] \subset \mathbb{R}^D$, the restriction space $C^\sigma(E)$, and the smooth-test metric d_σ defined in (43).

Fix $K \geq 3$. The affine chart

$$\mathscr{D}_K := \big\{(u, 1) : u \in \{0, 1, \dots, K\}^{D-1}\big\} \subset \mathbb{R}^D \tag{72}$$

contains $(K+1)^{D-1}$ directions. Choose constants $A_h, A_\tau \geq 1$, to be fixed below, and set

$$h_K := (K+1)^{-A_h}, \qquad \tau_K := (K+1)^{-A_\tau K}. \tag{73}$$

Using the probe from (178), define the normalized law features

$$\begin{aligned} \widehat{\mathcal{L}}^K_{a,j} &:= \mathcal{L}_{a/K, Kjh_K} = K^{-1} \mathcal{L}_{a, jh_K}, \\ \Psi_K(\mu) &:= \Big(\widehat{\mathcal{L}}^K_{a,j}(\mu)\Big)_{a \in \mathscr{D}_K,\ 0 \leq j < 2K} \in \mathbb{R}^{2K(K+1)^{D-1}}. \end{aligned} \tag{74}$$

With the fixed threshold for A_h chosen below, $0 \leq Kjh_K \leq 2K^2 h_K \leq 1$. Thus every normalized probe is an admissible scalar attention statistic with exponent parameter in $[0, 1]$. We augment these probes by the token–position endpoint coordinates retained unchanged by the residual connection:

$$\Phi_K(\mu, y) := \big(y, \Psi_K(\mu)\big) \in \mathbb{R}^{m_K}, \qquad m_K := D + 2K(K+1)^{D-1} \leq C_D K^D. \tag{75}$$

These coordinates are normalized log-Laplace derivatives. Their normalization preserves all information while keeping both the probe values and the attention matrices uniformly bounded in K. Unlike an unmasked construction with a token-dependent query, this causal construction needs no large bias to suppress query perturbations.

Two analytic ingredients connect the probes to metric resolution. The first recovers projected moments by high-order interpolation of the logarithmic derivative; the second approximates smooth test functions by polynomials while controlling their coefficients.

Lemma C.1 (Stable recovery of projected moments)**.** *Fix A_h above a threshold depending only on D, and let h_K be as in (73). There are constants $C_{\mathrm{rec}} = C_{\mathrm{rec}}(D, A_h) > 0$ and $C_{\mathrm{tr}} = C_{\mathrm{tr}}(D) > 0$ such that, for every $K \geq 3$, $a \in \mathscr{D}_K$, $\mu, \nu \in \mathcal{P}(E)$, and $1 \leq r \leq K$,*

$$\begin{aligned} &\left| \int_E \langle a, y \rangle^r \, \mathrm{d}(\mu - \nu)(y) \right| \\ &\qquad \leq (K+1)^{C_{\mathrm{rec}} K} \max_{0 \leq j < 2K} |\mathcal{L}_{a, jh_K}(\mu) - \mathcal{L}_{a, jh_K}(\nu)| + (K+1)^{C_{\mathrm{tr}} K} h_K^K. \end{aligned} \tag{76}$$

Proof. Analyticity and derivative bounds. Let $S := \langle a, Y \rangle$ for $Y \sim \mu$, and write $M_{\mu,a}(c) := \int_E e^{c\langle a, y \rangle} \, \mathrm{d}\mu(y)$ and $\ell_{\mu,a} := M'_{\mu,a} / M_{\mu,a}$. Uniformly over $a \in \mathscr{D}_K$ and $y \in E$, $|\langle a, y \rangle| \leq C_D K$. On a complex disk of radius c_D / K, $|M_{\mu,a}(c) - 1| \leq e^{C_D K |c|} - 1$; after decreasing c_D, this is at most $1/2$. Thus $M_{\mu,a}$ has no zero there and $\ell_{\mu,a}$ is analytic. On a smaller disk, $|\ell_{\mu,a}| \leq C_D K$; Cauchy's estimate therefore gives, for $0 \leq q \leq K$,

$$\sup_{0 \leq c \leq 2Kh_K} \left| \ell^{(q)}_{\mu,a}(c) \right| \leq q! \, (C_D K)^{q+1} \leq (K+1)^{C_D (q+1)}. \tag{77}$$

Here A_h is chosen once so that $[0, 2Kh_K]$ remains in a smaller zero-free disk for all $K \geq 3$. The same estimate holds for ν.

High-order interpolation. Put $N := 2K$, and let $I_{N,h_K}\ell$ be the polynomial of degree at most $N-1$ interpolating ℓ at $0, h_K, \dots, (N-1)h_K$. The cardinal polynomials on the integer grid are

$$\lambda_j(v) := \prod_{\substack{0 \leq m < N \\ m \neq j}} \frac{v-m}{j-m}, \qquad \sum_{j=0}^{N-1} \|\lambda_j\|_{\ell^1(\mathrm{coeff})} = 2^N - 1 \leq 2^N.$$

Indeed, the nonnegative roots give alternating coefficient signs, so $\|\lambda_j\|_{\ell^1(\mathrm{coeff})} = |\lambda_j(-1)| = \binom{N}{j+1}$; summing gives the identity. Since $I_{N,h_K}\ell(c) = \sum_{j=0}^{N-1} \ell(jh_K)\lambda_j(c/h_K)$, perturbing all samples by at most η changes its q-th derivative at zero by at most $q! h_K^{-q} 2^N \eta$.

Write $\ell(c) = \sum_{q \geq 0} t_q c^q$ near zero. The analytic disk bound gives $|t_q| \leq C_D K (C_D K)^q$ for every $q \geq 0$, not only the derivative orders through $K-1$ that we seek to recover. Choose the fixed threshold for A_h so that $C_D K N h_K \leq 1/2$ for every $K \geq 3$. Summing the geometric tail of the Taylor polynomial $T_{N-1}\ell$ then gives

$$\max_{0 \leq j < N} |\ell(jh_K) - T_{N-1}\ell(jh_K)| \leq C_D K (C_D K N h_K)^N.$$

The factor from summing the tail is absorbed into C_D. Interpolation is exact on $T_{N-1}\ell$, so for $0 \leq q < K$,

$$\left| \ell^{(q)}(0) - (I_{N,h_K}\ell)^{(q)}(0) \right| \leq q! h_K^{-q} 2^N C_D K (C_D K N h_K)^N \leq (K+1)^{C_2 K} h_K^K. \tag{78}$$

Here C_2 depends only on D: use $N = 2K$, $q < K$, and $h_K^{N-q} \leq h_K^K$. The $2K$ samples leave a full power h_K^K after differentiation through order $K-1$; a polynomially small spacing now suffices. This interpolation is used only to prove feature resolution, not as an additional layer of the transformer.

Set $\eta := \max_{0 \leq j < 2K} |\ell_{\mu,a}(jh_K) - \ell_{\nu,a}(jh_K)|$. The cumulants are $\kappa_j := \ell_{\mu,a}^{(j-1)}(0)$, and the preceding sample and bias estimates give, for $1 \leq j \leq r \leq K$,

$$|\kappa_j(\mu) - \kappa_j(\nu)| \leq (j-1)! h_K^{-(j-1)} 2^{2K} \eta + 2(K+1)^{C_2 K} h_K^K. \tag{79}$$

Because $h_K = (K+1)^{-A_h}$, the right-hand side is bounded by

$$(K+1)^{C_1 K} \eta + (K+1)^{C_2' K} h_K^K, \tag{80}$$

where C_1 may depend on A_h, while C_2' depends only on D. For $\lambda \in \{\mu, \nu\}$, write $\kappa(\lambda) := (\kappa_1(\lambda), \dots, \kappa_r(\lambda))$.

From cumulants to moments. Finally,

$$\int_E \langle a, y \rangle^r \, \mathrm{d}\mu(y) = B_r\big(\kappa_1, \dots, \kappa_r\big), \tag{81}$$

where B_r is the complete exponential Bell polynomial. The identity $\partial B_r / \partial \kappa_j = \binom{r}{j} B_{r-j}$ will control its gradient. More explicitly,

$$B_s(x_1, \dots, x_s) = \sum_{\substack{k_1, \dots, k_s \geq 0 \\ \sum_{j=1}^s j k_j = s}} s! \prod_{j=1}^s \frac{1}{k_j!} \left(\frac{x_j}{j!} \right)^{k_j}. \tag{82}$$

Equation (77) gives, after enlarging a dimension-dependent constant, $|\kappa_j| \leq j! (C_D K)^j$ for $j \leq K$, and the same bound holds along the segment between the two cumulant vectors. In each summand of (82), the powers of $C_D K$ total exactly s. The number of integer partitions, the factor $s!$, and the remaining combinatorial sum are all bounded by $(K+1)^{CK}$ for $s \leq K$. Using positivity of the coefficients and the derivative identity therefore gives

$$\sup_{0 \leq t \leq 1} \|\nabla B_r((1-t)\kappa(\mu) + t\kappa(\nu))\|_1 \leq (K+1)^{C_3 K}. \tag{83}$$

The mean-value theorem, together with (80), now gives (76) after enlarging C_{rec} and C_{tr}. □

Lemma C.2 (Jackson approximation with coefficient control)**.** *There is $C_0 > 1$, depending only on D, such that, for every fixed $\sigma > 0$, $K \geq 3$, and $g \in C^\sigma(E)$, there is a polynomial $p_K(y) = \sum_{|\kappa| \leq K} c_\kappa y^\kappa$ satisfying*

$$\|g - p_K\|_{L^\infty(E)} \leq C_\sigma K^{-\sigma} \|g\|_{C^\sigma(E)}, \qquad \sum_{|\kappa| \leq K} |c_\kappa| \leq C_\sigma C_0^K \|g\|_{C^\sigma(E)}. \tag{84}$$

Proof. Jackson approximation. If $g = 0$, take $p_K = 0$. Otherwise, choose an extension $\widetilde{g}$ to Q with $\|\widetilde{g}\|_{C^\sigma(\mathsf{Q})} \leq 2\|g\|_{C^\sigma(E)}$. After mapping the last coordinate affinely to $[-1, 1]$, apply bounded one-dimensional Jackson approximation operators of order exceeding the fixed σ, successively in all D coordinates, with coordinate degree $\lfloor K/D \rfloor$. Their bounds may depend on σ; a fixed low-order positive operator would not suffice at arbitrary smoothness. This gives the standard tensor multivariate Jackson estimate (DeVore & Lorentz, 1993; Totik, 2020). It provides a polynomial of total degree at most K with the first stated error above a threshold depending on (D, σ). The finitely many smaller $K \geq 3$ are handled by a constant polynomial after enlarging C_σ. In all cases,

$$\|p_K\|_{L^\infty(\mathsf{Q})} \leq \|\widetilde{g}\|_{L^\infty(\mathsf{Q})} + \|\widetilde{g} - p_K\|_{L^\infty(\mathsf{Q})} \leq C_\sigma \|g\|_{C^\sigma(E)}.$$

Coefficient control. Write $m := \lfloor K/D \rfloor$. The tensor Jackson construction has coordinate degree at most m, so after the affine change in the last coordinate it has an expansion

$$p_K(y) = \sum_{j \in \{0,\dots,m\}^D} \widehat{c}_j \prod_{\ell=1}^{D} T_{j_\ell}(\widetilde{y}_\ell),$$

where $\widetilde{y}_D = 2y_D - 1$ and the other coordinates are unchanged. The tensor Chebyshev integral formula gives $|\widehat{c}_j| \leq 2^D \|p_K\|_{L^\infty(\mathsf{Q})}$. The sum of the absolute monomial coefficients of a degree-s Chebyshev polynomial, including after the affine change in the last coordinate, is at most C^s. Hence each product above has monomial coefficient norm at most $C^{|j|} \leq C^K$, and it has total degree $|j| \leq Dm \leq K$. There are $(m+1)^D \leq (K+1)^D$ tensor coefficients. Since D is fixed and $K \geq 3$, this polynomial factor is absorbed into C_0^K, with C_0 independent of σ. This proves the second inequality. □

Proposition C.3 (Quantitative resolution by exact causal probes)**.** *For every prescribed $M_0 > 0$, constants $A_h = A_h(D, M_0)$ and $A_\tau = A_\tau(D, M_0)$ can be chosen so that, for all $K \geq 3$, $\mu, \nu \in \mathcal{P}(E)$, and $y, y' \in E$,*

$$\begin{aligned} &\|\Phi_K(\mu, y) - \Phi_K(\nu, y')\|_\infty \leq \tau_K \\ \Longrightarrow &\max_{1 \leq |\kappa| \leq K} \left| \int_E w^\kappa \, \mathrm{d}(\mu - \nu)(w) \right| \leq (K+1)^{-M_0 K}. \end{aligned} \tag{85}$$

There is a threshold $M_\star = M_\star(D)$ such that, whenever $M_0 \geq M_\star$, this implies

$$D_{\mathrm{prod}}\big((\mu, y), (\nu, y')\big) \leq CK^{-1}, \tag{86}$$

$$d_\sigma(\mu, \nu) + \|y - y'\|_\infty \leq C_\sigma K^{-\sigma} \qquad (\sigma > 0 \textit{ fixed}). \tag{87}$$

After fixing one $M_0 \geq M_\star$, the same feature family and the same h_K, τ_K work for every fixed σ; only C_σ depends on σ.

Proof. We propagate feature closeness in three stages: normalized probes control projected moments, interpolation recovers mixed moments, and Jackson approximation controls smooth tests.

Projected moments. Put $\Lambda_D := 4^{D-1}$, and choose an integer $C_{\mathrm{int}}(D) \geq 0$ such that, for every $K \geq 3$,

$$2\Lambda_D^K (K+1)^{-(C_{\mathrm{int}}(D)+1)K} \leq 1, \tag{88}$$

and set $M := M_0 + C_{\mathrm{int}}(D) + 1$. First choose $A_h \geq M + C_{\mathrm{tr}}(D) + 1$ and above the threshold in Lemma C.1. The truncation term is then at most $(K+1)^{-MK}$. With this A_h fixed, choose

$A_\tau \geq C_{\mathrm{rec}}(D, A_h) + M + 1$. The interpolation amplification will then absorb both τ_K and the normalization below. If the two normalized feature vectors are τ_K-close, (74) gives

$$\max_{a \in \mathscr{D}_K,\ 0 \leq j < 2K} |\mathcal{L}_{a,jh_K}(\mu) - \mathcal{L}_{a,jh_K}(\nu)| \leq K\tau_K.$$

The additional factor K is absorbed by the extra unit in the choice of A_τ, since

$$(K+1)^{C_{\mathrm{rec}}K} K\tau_K \leq (K+1)^{1-(M+1)K} \leq (K+1)^{-MK}.$$

Lemma C.1 therefore bounds every projected moment through degree K by $2(K+1)^{-MK}$.

Mixed moments. For fixed r, the projected moment is the homogeneous polynomial

$$a \longmapsto \int \langle a, y\rangle^r \, \mathrm{d}(\mu - \nu)(y) = \sum_{|\kappa|=r} \binom{r}{\kappa} a^\kappa \int y^\kappa \, \mathrm{d}(\mu - \nu)(y). \tag{89}$$

Set $P_r(a) \coloneqq \int \langle a, y\rangle^r \, \mathrm{d}(\mu - \nu)(y)$ and $Q_r(u) \coloneqq P_r((u, 1))$. The polynomial Q_r has total degree at most r, and tensor-product Lagrange interpolation on $u \in \{0, \dots, r\}^{D-1}$ gives the exact identity

$$Q_r(u) = \sum_{j \in \{0,\dots,r\}^{D-1}} Q_r(j) \prod_{\ell=1}^{D-1} \lambda_{j_\ell}(u_\ell).$$

The required grid values are available because $r \leq K$. In one variable the cardinal basis is

$$\lambda_j(v) \coloneqq \prod_{\substack{0 \leq m \leq r \\ m \neq j}} \frac{v - m}{j - m}, \qquad \|\lambda_j\|_{\ell^1(\mathrm{coeff})} = \binom{r+1}{j+1}, \qquad \sum_{j=0}^{r} \|\lambda_j\|_{\ell^1(\mathrm{coeff})} \leq 2^{r+1}.$$

Consequently,

$$\|Q_r\|_{\ell^1(\mathrm{coeff})} \leq \max_{j \in \{0,\dots,r\}^{D-1}} |Q_r(j)| \left(\sum_{j=0}^{r} \|\lambda_j\|_{\ell^1(\mathrm{coeff})} \right)^{D-1} \leq \Lambda_D^r \max_j |Q_r(j)| \,.$$

For every $\kappa' \in \mathbb{N}_0^{D-1}$ with $|\kappa'| \leq r$, the coefficient of $u^{\kappa'}$ is $\binom{r}{\kappa', r-|\kappa'|} \int y^{(\kappa', r-|\kappa'|)} \, \mathrm{d}(\mu - \nu)(y)$, while coefficients with $|\kappa'| > r$ vanish. Thus every degree-r mixed moment is recovered. Division by its multinomial coefficient cannot enlarge the bound. Multiplying the projected-moment estimate by Λ_D^K and using (88) gives the right-hand side of (85). This proves (85).

Smooth-test and Wasserstein resolution. Apply Lemma C.2 to a test function $g \in C^\sigma(E)$. Integrating its polynomial approximation against $\mu - \nu$, and using (85), yields

$$\left| \int_E g \, \mathrm{d}(\mu - \nu) \right| \leq C_\sigma \, \|g\|_{C^\sigma(E)} \left(K^{-\sigma} + C_0^K (K+1)^{-M_0 K} \right). \tag{90}$$

The constant monomial cancels because μ and ν have equal total mass. Take $M_\star = 1$. For every fixed $\sigma > 0$ and every $M_0 \geq M_\star$, $C_0^K (K+1)^{-M_0 K} \leq C_\sigma K^{-\sigma}$ uniformly for $K \geq 3$; importantly, the feature scales do not depend on σ. This proves (87). For (86), fix $y_0 \in E$, subtract the value at y_0 from each unit-Lipschitz Kantorovich potential, and take its McShane extension to Q, using the metric in (70). Its $C^{0,1}(\mathsf{Q})$ norm is bounded uniformly in the potential, up to fixed norm-equivalence constants. Kantorovich–Rubinstein duality gives the claim. In both cases, the retained coordinate block gives $\|y - y'\|_\infty \leq \tau_K$, which is smaller than every fixed power of K^{-1}. □

C.2 A FIRST-VARIATION CERTIFICATE FOR HIGH-ORDER REGULARITY

For $\beta > 1$, the theorem combines order-β smooth-test stability in the history law with first-order stability in the retained endpoint coordinates. Here $C^\beta(E)$ means joint regularity in token and normalized position; it is not merely token regularity or an unspecified notion of functional smoothness. The next lemma turns a familiar first-variation assumption into exactly the seminorm required by the theorem.

Lemma C.4 (Smooth first variation implies the β-smooth target condition)**.** *Fix $\beta > 1$. Let $\mathbf{F}^\star$ be a completed causal family with canonical factor $f^\star : \mathcal{M} \to \mathbb{R}^{d'}$. Suppose there is a map $\widetilde{f} : \mathcal{P}(E) \times E \to \mathbb{R}^{d'}$ satisfying*

$$\widetilde{f}(\mu, \mathfrak{e}(\mu)) = f^\star(\mu), \qquad \mu \in \mathcal{M}, \tag{91}$$

and, for some $B_\beta \geq 0$,

$$\sup_{\mu \in \mathcal{P}(E)} \sup_{y \neq y'} \frac{\left\| \widetilde{f}(\mu, y) - \widetilde{f}(\mu, y') \right\|}{\|y - y'\|_\infty} \leq B_\beta. \tag{92}$$

Assume each coordinate has a jointly Borel measurable linear functional derivative, with all displayed integrals well-defined, such that for $\mu_u = (1-u)\mu + u\nu$,

$$\widetilde{f}_\ell(\nu, y) - \widetilde{f}_\ell(\mu, y) = \int_0^1 \int_E \frac{\delta \widetilde{f}_\ell}{\delta \mu}(\mu_u, y, v) \, \mathrm{d}(\nu - \mu)(v) \, \mathrm{d}u. \tag{93}$$

Define

$$\|\varphi\|_{\dot{C}^\beta(E)} := \inf_{c \in \mathbb{R}} \|\varphi + c\|_{C^\beta(E)}, \tag{94}$$

and suppose

$$\left(\sum_{\ell=1}^{d'} \sup_{(\mu, y) \in \mathcal{P}(E) \times E} \left\| \frac{\delta \widetilde{f}_\ell}{\delta \mu}(\mu, y, \cdot) \right\|_{\dot{C}^\beta(E)}^2 \right)^{1/2} \leq B_\beta. \tag{95}$$

Then $R_{\beta;\omega}(\mathbf{F}^\star) \leq B_\beta$.

Proof. Quotient duality. For every $\varphi \in C^\beta(E)$,

$$\left| \int_E \varphi \, \mathrm{d}(\nu - \mu) \right| \leq \|\varphi\|_{\dot{C}^\beta(E)} \, d_\beta(\mu, \nu), \tag{96}$$

because additive constants integrate to zero. Thus $\dot{C}^\beta(E)$ is the relevant quotient seminorm; no converse duality identity is needed. For $\mu, \nu \in \mathcal{M}$, insert $\widetilde{f}(\nu, \mathfrak{e}(\mu))$ between $\widetilde{f}(\nu, \mathfrak{e}(\nu))$ and $\widetilde{f}(\mu, \mathfrak{e}(\mu))$. The first-variation identity, the Euclidean aggregate in (95), and (96) control the measure part by $B_\beta d_\beta(\mu, \nu)$, while (92) controls the remaining part by $B_\beta \|\mathfrak{e}(\mu) - \mathfrak{e}(\nu)\|_\infty$. Using the trace identity (91) and Definition B.2 proves the claim. □

The ambient formulation makes the certificate easy to state and verify, but the proof only uses (93) and (95) along mixture segments $(1-u)\mu + u\nu$ joining $\mu, \nu \in \mathcal{M}_\omega$. Likewise, the endpoint estimate is needed only at the states appearing in the preceding split.

C.3 A DIMENSION-EXPLICIT SHALLOW RELU READOUT

The feature dimension grows with the resolution scale, so a qualitative universal-approximation theorem is insufficient for counting parameters. The next lemma makes the dimension dependence explicit while retaining one hidden ReLU layer.

Lemma C.5 (Counted shallow approximation of a Lipschitz map)**.** *Fix $B_0 \geq 1$. There is $C_{B_0} > 0$ such that the following holds for every $m \geq 1$. Let $g : [-B_0, B_0]^m \to \mathbb{R}$ satisfy*

$$\|g\|_\infty \leq M, \qquad |g(u) - g(v)| \leq A \|u - v\|_\infty, \tag{97}$$

with $A, M \geq 0$. If $M = 0$, the zero network is exact. If $M > 0$, then for every $0 < \delta \leq M$, there is a standard one-hidden-layer ReLU network $\mathcal{R}$ of width W such that

$$\|g - \mathcal{R}\|_\infty \leq \delta \tag{98}$$

and

$$\log(2 + W) \leq C_{B_0} \left[(m+2) \log\left(2 + \frac{A}{\delta}\right) + m + \log\left(2 + \frac{M}{\delta}\right) + \log(m+1) \right]. \tag{99}$$

When $A = 0$, a constant network gives the conclusion.

Proof. It remains to consider $A, M > 0$. *Extension and mollification.* Let $\widetilde{g}$ be the McShane extension of g to $\mathbb{R}^m$, clipped to $[-M, M]$. It still agrees with g on the cube and is A-Lipschitz for $\|\cdot\|_\infty$. Choose a smooth tensor-product cutoff that equals one on $[-B_0 - 1, B_0 + 1]^m$, is supported in $[-B_0 - 2, B_0 + 2]^m$, and multiply $\widetilde{g}$ by it to obtain a compactly supported function G. Then

$$\|G\|_{L^1(\mathbb{R}^m)} \le M C_0^m \tag{100}$$

for a constant $C_0 = C_0(B_0)$.

Let $\zeta \in C_c^\infty([-1, 1])$ be nonnegative with integral one, put $\zeta_m := \zeta^{\otimes m}$, and mollify at scale

$$h := \min\left\{1, \frac{\delta}{3A}\right\}, \qquad G_h := G * \zeta_{m,h}. \tag{101}$$

On $[-B_0, B_0]^m$, the convolution only sees the region where the cutoff is one. Hence

$$\|G_h - g\|_\infty \le Ah \le \delta/3. \tag{102}$$

Fourier complexity. The mollified function G_h lies in $C_c^\infty(\mathbb{R}^m)$, so it has the Fourier representation required below. For a fixed Fourier-transform convention, define $v_{G_h,2} := \int_{\mathbb{R}^m} \|\omega\|_1^2 \left|\widehat{G}_h(\omega)\right| \, \mathrm{d}\omega$. Since $\left|\widehat{G}\right| \le C \|G\|_{L^1}$, a change of variables and the tensor structure give

$$v_{G_h,2} \le C M m^2 C_1^m h^{-m-2}, \tag{103}$$

where $C_1 = C_1(B_0, \zeta) \ge 3$; indeed, $\int \|\xi\|_1^2 \prod_{j=1}^m \left|\widehat{\zeta}(\xi_j)\right| \, \mathrm{d}\xi \le m^2 C_1^m$. All Fourier-normalization constants are absorbed into C_1^m.

Counted ridge approximation. To record the rescaling explicitly, set $G_h^{(B_0)}(x) := G_h(B_0 x)$. This function is smooth and compactly supported, so Fourier inversion applies, and a change of variables gives $v_{G_h^{(B_0)},2} = B_0^2 v_{G_h,2}$ up to the fixed Fourier convention. The factor B_0^2 is harmless because B_0 is fixed. Theorem 2 of Klusowski & Barron (2018), applied on $[-1, 1]^m$ and then composed with the inverse scaling, supplies a one-hidden-layer ReLU ridge network with W units and an affine term whose uniform error on $[-B_0, B_0]^m$ is at most

$$C v_{G_h,2} \sqrt{m + \log W} \, W^{-1/2-1/m}. \tag{104}$$

Put $x := v_{G_h,2}/\delta$ and take $W := \left\lceil [C_*(m+1)(1+x)]^4 \right\rceil$, where $C_* = C_*(B_0)$ is sufficiently large and independent of m, g, A, M, δ. Since $W^{-1/m} \le 1$, substitution in (104) gives

$$\frac{C v_{G_h,2} \sqrt{m + \log W} \, W^{-1/2-1/m}}{\delta} \le \frac{C x \sqrt{m + \log W}}{C_*^2 (m+1)^2 (1+x)^2}. \tag{105}$$

The ceiling obeys

$$\log W \le \log 2 + 4 \log C_* + 4 \log(m+1) + 4 \log(1+x).$$

Consequently the right-hand side of (105) is bounded by a constant times

$$C_*^{-2} \frac{x}{(1+x)^2} \frac{\sqrt{m + \log C_* + \log(m+1) + \log(1+x)}}{(m+1)^2}.$$

This tends to zero uniformly over $m \ge 1$ and $x \ge 0$ as $C_* \to \infty$: besides the elementary polynomial bounds, one uses $\sup_{x \ge 0} x(1+x)^{-2} \sqrt{\log(1+x)} < \infty$. Choose C_* so that the bound is at most $2/3$. Thus (104) is at most $2\delta/3$. Combining this with (102) proves (98). If the pointwise MLP has no direct affine skip, its affine term is represented by the additional units $u_j = \mathrm{ReLU}(u_j) - \mathrm{ReLU}(-u_j)$, costing at most $2m$ units. Replace W by $W + 2m$ and relabel; the logarithmic bound is unchanged. Finally, (103) and $h^{-1} \le 1 + 3A/\delta$ give

$$\log(1+x) \le C_{B_0} + \log\left(2 + \frac{M}{\delta}\right) + 2\log(m+1) + m \log C_1 + (m+2) \log\left(2 + \frac{A}{\delta}\right).$$

Combining this inequality with the definition of W, and absorbing fixed and lower-order terms, proves (99) after enlarging C_{B_0}. □

C.4 Quantitative readout and total parameter count

We now assemble the proof. Feature resolution first controls the target oscillation on each feature cell; a shallow readout then approximates the resulting feature-space map; finally, we count the full construction. The same parameters are shared across all sequence lengths, and the final budget includes every attention and readout coefficient.

For the remainder of the proof, take $M_0 = 1$ in Proposition C.3, as permitted by its proof, and fix the corresponding A_h, A_τ. These scale constants depend only on D and work for the $\sigma = \beta_+$ smooth-test estimate; in particular, they do not depend on ω or the target.

Target resolution in feature space. Let $\mu, \nu \in \mathcal{M}$, set $y := \mathfrak{e}(\mu)$ and $y' := \mathfrak{e}(\nu)$, and suppose that $\Phi_K(\mu, y)$ and $\Phi_K(\nu, y')$ are τ_K-close in the supremum norm. Applying (87) with $\sigma = \beta_+$, and using the retained endpoint coordinates, gives

$$\Delta_\beta(\mu,\nu) = d_{\beta_+}(\mu,\nu)^{\beta_-} + \|y - y'\|_\infty^{\beta_-} \le C_\beta K^{-\beta_+\beta_-} + \tau_K^{\beta_-} \le C_\beta K^{-\beta}.$$

Here $\beta_+\beta_- = \beta$, and τ_K decays faster than every fixed inverse power of K. Definition B.2 yields

$$\|\Phi_K(\mu,\mathfrak{e}(\mu)) - \Phi_K(\nu,\mathfrak{e}(\nu))\|_\infty \le \tau_K \quad\Longrightarrow\quad \|f^\star(\mu) - f^\star(\nu)\| \le C_\beta R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}. \tag{106}$$

This is the decisive regularity step: a feature cell of radius τ_K has target oscillation $O(K^{-\beta})$. The full power $K^{-\beta}$ is therefore obtained before introducing the first-order Lipschitz envelope and its shallow ReLU approximation; neither construction causes saturation at $\beta = 1$.

Bounded feature range and scalar reduction. Every coordinate of Ψ_K is a weighted average of $\langle a/K, y\rangle$, while the retained endpoint coordinates are bounded on E. Thus there is a fixed $B_0 = B_0(D) \ge 1$ such that

$$\Phi_K(\mathcal{P}(E) \times E) \subset [-B_0, B_0]^{m_K} \qquad \text{for every } K \ge 3. \tag{107}$$

The seminorm $R_{\beta;\omega}$ controls oscillations but not an absolute offset. We first separate these two roles. There is $C_{\mathrm{osc}} = C_{\mathrm{osc}}(D) \ge 1$ such that

$$\sup_{\mu,\nu\in\mathcal{M}} \|f^\star(\mu) - f^\star(\nu)\| \le C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star). \tag{108}$$

Indeed, $d_{\beta_+} \le 2$, because the full test-function norm controls the uniform norm, while the endpoint diameter is at most two. Since $0 < \beta_- \le 1$, the diameter for Δ_β is therefore at most four. Fixing $\mu_0 \in \mathcal{M}$, center each output coordinate by

$$g_\ell(\mu) := f_\ell^\star(\mu) - f_\ell^\star(\mu_0). \tag{109}$$

For the moment, fix ℓ, write $g := g_\ell$, and suppress the coordinate index. Then

$$\|g\|_{L^\infty(\mathcal{M})} \le C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star). \tag{110}$$

Lipschitz envelope. Assume first that $R_{\beta;\omega}(\mathbf{F}^\star) > 0$. On $\mathsf{Q}_K := [-B_0, B_0]^{m_K}$, define

$$\begin{aligned} R_K(u) &:= \inf_{\mu\in\mathcal{M}} \left\{ g(\mu) + A_K \left\| u - \Phi_K(\mu,\mathfrak{e}(\mu)) \right\|_\infty \right\}, \\ A_K &:= C_A R_{\beta;\omega}(\mathbf{F}^\star)/\tau_K, \qquad C_A > C_{\mathrm{osc}}. \end{aligned} \tag{111}$$

For $\lambda \in \mathcal{M}$, abbreviate $u_\lambda := \Phi_K(\lambda, \mathfrak{e}(\lambda))$. We verify the envelope on this feature image. Fix $\lambda \in \mathcal{M}$. Using λ itself as a competitor gives $R_K(u_\lambda) \le g(\lambda)$. Conversely, a competitor ν with $\|u_\lambda - u_\nu\|_\infty \le \tau_K$ has value at least $g(\lambda) - C_\beta R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}$ by (106). If its feature distance instead exceeds τ_K, the global oscillation bound and $C_A > C_{\mathrm{osc}}$ give

$$g(\nu) + A_K \|u_\lambda - u_\nu\|_\infty \ge g(\lambda) - C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star) + C_A R_{\beta;\omega}(\mathbf{F}^\star) \ge g(\lambda). \tag{112}$$

Every competitor falls into one of these cases, so taking the infimum proves

$$\sup_{\mu\in\mathcal{M}} |g(\mu) - R_K(\Phi_K(\mu,\mathfrak{e}(\mu)))| \le C_\beta R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}. \tag{113}$$

The envelope is A_K-Lipschitz. Project it onto the interval containing the centered target range:

$$\overline{R}_K := \mathrm{proj}_{[-C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star), C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star)]} \circ R_K. \tag{114}$$

This projection is one-Lipschitz and fixes every centered target value, so (113) remains valid with $\overline{R}_K$, while

$$\left\|\overline{R}_K\right\|_\infty \le C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star), \qquad \mathrm{Lip}_\infty(\overline{R}_K) \le A_K. \tag{115}$$

Counted shallow readout. Repeat the envelope construction for each coordinate ℓ, using the same feature map and constants. Apply Lemma C.5 coordinatewise with $M = C_{\mathrm{osc}} R_{\beta;\omega}(\mathbf{F}^\star)$, $A = A_K$, and $\delta := R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}$. The shallow-network estimate uses

$$\log\left(2 + \frac{A_K}{\delta}\right) \le C_\beta + A_\tau K \log(K+1) + \beta \log K. \tag{116}$$

The scale $R_{\beta;\omega}(\mathbf{F}^\star)$ cancels from A_K/δ; hence neither this ratio nor the feature dimension depends on ω. Since $m_K \le C_D K^D$, concatenate the hidden units of the d' scalar networks and stack their output rows to obtain one pointwise network. The widths add, while coordinatewise error δ produces Euclidean error at most $\sqrt{d'}\,\delta$, absorbed into $C_\beta = C(d, d', \beta)$. Its total hidden width W_K satisfies

$$\log(2 + W_K) \le C_\beta K^{D+1} \log(K+1) = C_\beta K^{d+2} \log(K+1). \tag{117}$$

This exponent has a direct origin. The joint token–time variable has dimension $D = d + 1$, giving $m_K = O(K^D)$ features; each feature contributes the $O(K \log(K+1))$ logarithmic stability cost in (116). Their product yields $K^{d+2} \log(K+1)$, hence the denominator $d + 2 = (d+1) + 1$ in the final rate. Adding the base values $f^\star_\ell(\mu_0)$ through the final affine bias gives

$$\sup_{\mu \in \mathcal{M}} \|f^\star(\mu) - \mathcal{R}_K(\Phi_K(\mu, \mathfrak{e}(\mu)))\| \le C_\beta R_{\beta;\omega}(\mathbf{F}^\star) K^{-\beta}. \tag{118}$$

Exact attention realization. It remains to realize the abstract feature map inside the transformer. Each of its $H_K := 2K(K+1)^{D-1} = O(K^D)$ normalized probe coordinates is computed exactly by one masked head. With the homogeneous and work coordinates appended, take

$$\begin{aligned} \mathsf{Q}^{\mathrm{att}}_{a,j} &:= (0_D, 1, 0_{H_K}), & \mathsf{K}^{\mathrm{att}}_{a,j} &:= (j h_K a^\top, 0, 0_{H_K}), \\ \mathsf{V}^{\mathrm{att}}_{a,j} &:= ((a/K)^\top, 0, 0_{H_K}), & \mathsf{W}^{\mathrm{att}}_{a,j} &:= e_{D+1+r(a,j)}, \end{aligned} \tag{119}$$

where $r(a,j) \in [H_K]$ enumerates the probes. These matrices realize $\mathcal{L}_{a/K, K j h_K} = K^{-1} \mathcal{L}_{a, j h_K}$ on every finite prefix and on the temporal boundary. Every displayed entry lies in $[-1, 1]$: the query and output matrices have only zero–one entries, $\|a/K\|_\infty \le 1$, and $\|j h_K a\|_\infty \le 2K^2 h_K \le 1$.

Total parameter count. Let $r_K := D + 1 + H_K = d + 2 + H_K$ be the embedding width. Thus every matrix $\mathsf{Q}^{\mathrm{att}}_{a,j}$, $\mathsf{K}^{\mathrm{att}}_{a,j}$, and $\mathsf{V}^{\mathrm{att}}_{a,j}$ lies in $\mathbb{R}^{1 \times r_K}$, while $\mathsf{W}^{\mathrm{att}}_{a,j} \in \mathbb{R}^{r_K \times 1}$. The combined readout $\mathcal{R}_K$ was constructed on the $m_K = D + H_K$ coordinates (y, Ψ_K). Identify it on the actual transformer state by

$$\widehat{\mathcal{R}}_K(y, 1, v) := \mathcal{R}_K(y, v), \tag{120}$$

which inserts one zero column, at the homogeneous-coordinate position, in the first readout matrix. The resulting width-W_K pointwise readout has the form (5), with

$$A_{1,K} \in \mathbb{R}^{W_K \times r_K}, \quad b_{1,K} \in \mathbb{R}^{W_K}, \qquad A_{2,K} \in \mathbb{R}^{d' \times W_K}, \quad b_{2,K} \in \mathbb{R}^{d'}.$$

We now count the architecture just constructed, rather than an abstract feature budget. Let T_K denote the resulting shallow transformer. Counting every scalar slot in the attention parameter list θ_K and the readout parameter list η_K gives the exact dense count

$$\begin{aligned} p_K := \mathrm{par}(T_K) &= 4 H_K r_K + (r_K + d' + 1) W_K + d', \\ \log p_K &\le C_\beta K^{d+2} \log(K+1). \end{aligned} \tag{121}$$

Here $4 H_K r_K$ counts the four length-r_K matrices in every scalar head; the remaining terms count both readout matrices and both biases. Thus p_K includes structural zeros and all biases, while the fixed coordinate lift contributes no parameters and shared coefficients are counted once. As stated in Theorem B.3, this dense real-parameter count does not itself charge coefficient magnitudes or numerical precision. The normalized finite-precision refinement below controls these additional resources separately.

Inverting the count. Fix the final constant C_β in the parameter bound, enlarging it once and for all if necessary. Then enlarge $p_0 = p_0(d, d', \beta)$ so that it is at least 16, covers the fixed constant-output architecture described below, and satisfies

$$\exp\bigl(C_\beta 3^{d+2} \log 4\bigr) \le p_0.$$

Then for every $p \ge p_0$ the set in the following definition is nonempty:

$$K(p) := \max\left\{K \in \mathbb{N} :\ K \ge 3, \exp\bigl(C_\beta K^{d+2} \log(K+1)\bigr) \le p\right\}. \tag{122}$$

We write $T_p := T_{K(p)}$. Inverting this monotone budget relation gives constants $c_\beta, C'_\beta > 0$ such that

$$c_\beta \left(\frac{\log p}{\log\log p}\right)^{1/(d+2)} \le K(p) \le C'_\beta \left(\frac{\log p}{\log\log p}\right)^{1/(d+2)}. \tag{123}$$

Indeed, the defining inequality bounds $K(p)$ from above, whereas its failure at $K(p)+1$ gives the matching lower bound. Equations (118)– (123) give the discrete error bound at the rate in (52) and the head bound (53). The independent transfer identity (55) gives the stated joint error bound with the same parameters. For $0 < \varepsilon \le R_{\beta;\omega}(\mathbf{F}^\star)$, choosing

$$K_\varepsilon = \max\left\{3, \left\lceil \bigl(C_\beta R_{\beta;\omega}(\mathbf{F}^\star)/\varepsilon\bigr)^{1/\beta} \right\rceil\right\} \tag{124}$$

and substituting into (121) gives (54). When $R_{\beta;\omega}(\mathbf{F}^\star) = 0$, the canonical factor is constant on $\mathcal{M}$. To respect the theorem's conventions $H, M \ge 1$, take one dummy head with zero output column, a one-unit readout with zero output weight, and set its final bias equal to that constant. This realizes the target exactly and uses only the fixed number of parameters already absorbed into p_0.

Proof of Theorem B.3. Proposition C.3 converts proximity of exact attention features into proximity in the target metric, and Definition B.2 then gives (106). The clipped Lipschitz envelope and Lemma C.5 yield (118) at the width in (117). The heads in (119) realize these features at every finite resolution and on the temporal boundary, and (121) counts the resulting transformer. Inverting this count proves the discrete error bound, the head bound (53), and the accuracy-to-parameter estimate (54). The transfer identity (55), whose proof uses only continuous extendability and Proposition F.2, upgrades these estimates to the joint conclusion (52) without additional parameters or loss. □

C.5 Finite precision on the normalized target ball

The real-parameter construction also yields a finite binary description after normalizing the target amplitude. We track its coefficients and their sensitivity to rounding, uniformly over finite and continuous prefixes.

Proof of Corollary B.5. We first bound the coefficients of a common scale-K architecture, then quantize its dense parameter list. The estimates concern parameter storage; the mathematical network still evaluates its operations exactly.

Coefficient bounds and a common architecture. Write $R := R_{\beta;\omega}(\mathbf{F}^\star) \le 1$. If $R > 0$, the scalar envelopes above have magnitude at most $C_{\mathrm{osc}} R$, Lipschitz constant $A_K = C_A R/\tau_K$, and readout tolerance $\delta = RK^{-\beta}$. Their mollification scale in (101) is therefore

$$h_K^{\mathrm{rd}} := \frac{K^{-\beta}\tau_K}{3C_A} < 1.$$

For each scalar mollified envelope, (103) gives

$$v_{G_h,2} \le CRm_K^2 C_1^{m_K} (h_K^{\mathrm{rd}})^{-m_K-2}, \qquad \log\left(2 + \frac{v_{G_h,2}}{R}\right) \le C_\beta K^{D+1} \log(K+1). \tag{125}$$

Theorem 2 of Klusowski & Barron (2018), used in (104), supplies ridge directions of ℓ^1 norm one, thresholds in $[0,1]$, and coefficients of the form vb_k/W, where $|b_k| \le 1$ and $v \le C_{B_0} v_{G_h,2}$. Composing with $u \mapsto u/B_0$ keeps each inner row's ℓ^1 norm and each bias bounded by one. Its affine part has constant coefficient $G_h(0)$ and slope $\nabla G_h(0)$. Since mollification only sees the uncut Lipschitz extension near zero,

$$|G_h(0)| \le C_{\mathrm{osc}} R, \qquad \|\nabla G_h(0)\|_1 \le A_K.$$

Representing this affine part by the $2m_K$ additional ReLU units from the proof of Lemma C.5 preserves the inner-row and bias bounds. Its scalar output row has ℓ^1 norm at most $v + 2A_K$. Concatenating coordinates and inserting the zero homogeneous column preserve these row bounds. The restored target offset is bounded by one. These estimates bound every coefficient and give the more useful row-norm bounds

$$\begin{aligned} \max_i \sum_j |(A_{1,K})_{ij}| \le 1, \qquad \|b_{1,K}\|_\infty \le 1, \\ \max_\ell \sum_i |(A_{2,K})_{\ell i}| \le \exp(C_\beta K^{D+1} \log(K+1)). \end{aligned} \tag{126}$$

All scale choices are independent of the target. In particular, $v_{G_h,2}/\delta$ has a uniform upper bound by (125). Choose the scalar widths using this bound in (105), or pad smaller readouts with zero-output units. Concatenating the coordinates then gives a single topology for the entire normalized ball. It also includes the case $R = 0$: zero the readout weights and use its final bias for the constant target. After enlarging the design constant, its dense parameter count, every coefficient magnitude, and every output-row ℓ^1 norm are at most

$$P_K := \exp\big(C_\beta K^{D+1} \log(K+1)\big) .$$

Choose $K = K(p)$ by the budget inversion with this enlarged constant. Then these three bounds are at most p, while (126) retains the unit inner-row and bias bounds. The unrounded joint approximation error is at most $C_\beta K^{-\beta}$.

Uniform perturbation of dense parameters. Perturb every scalar parameter by at most $\eta \le 1/p$, retaining the same topology. Each input to the attention block is $(y, 1, 0_{H_K})$, with only $D+1$ possibly nonzero entries bounded by one. Since the original attention entries lie in $[-1, 1]$, scalar queries, keys, values, and their perturbed versions are bounded by a constant depending only on D. Their differences, and the differences of the scalar scores, are at most $C_D\eta$.

For any probability law λ and bounded real score functions s, s', put $\pi_s := e^s\lambda / \int e^s \, \mathrm{d}\lambda$. Differentiating along the segment joining the scores gives the dimension-free estimate

$$\int |\mathrm{d}\pi_s - \mathrm{d}\pi_{s'}| \le 2 \|s - s'\|_\infty .$$

Consequently each normalized head average changes by at most $C_D\eta$. The output column is also perturbed entrywise by at most η, so the same bound holds for each coordinate of a head's contribution. Summing the heads gives, for the exact and perturbed attention states u, u',

$$\|u - u'\|_\infty \le C_D H_K \eta, \qquad \|u\|_\infty + \|u'\|_\infty \le C_D. \tag{127}$$

The second estimate uses $H_K \le p$ and the bounded original feature range. This argument permits perturbations of structural zeros, including the dense output columns. It applies to every discrete prefix and to the normalized prefix integral for every $t > 0$, with no dependence on prefix length. At $t = 0$, the head average is simply the value at zero and the same bound follows directly.

Let r and M be the residual and readout widths; both are at most p. Write $v := A_1 u + b_1$ and $v' := A_1' u' + b_1'$. The unit inner-row bound and arbitrary dense entrywise perturbations give

$$\|v - v'\|_\infty \le \|u - u'\|_\infty + r\eta \|u'\|_\infty + \eta \le C_D p\eta, \qquad \|v'\|_\infty \le C_D.$$

ReLU is one-Lipschitz, and each original output row has ℓ^1 norm at most p. Perturbing the second affine layer therefore gives

$$\|A_2\rho(v) + b_2 - A_2'\rho(v') - b_2'\|_\infty \le p \|v - v'\|_\infty + M\eta \|\rho(v')\|_\infty + \eta.$$

Absorbing the fixed output-norm conversion, we obtain

$$\sup_{\text{all finite and boundary states}} \|T - T'\| \le C_{d,d'} p^2 \eta. \tag{128}$$

No sparsity is assumed in these readout perturbations.

Dyadic rounding and binary descriptions. For a sufficiently large constant C, choose

$$b := \lceil \log_2(Cp^2 K^\beta) \rceil , \qquad \eta := 2^{-b} \le \frac{K^{-\beta}}{Cp^2},$$

and round each parameter to the nearest multiple of 2^{-b}. The rounding error is at most $\eta \le 1/p$, hence (128) adds at most $K^{-\beta}$ to the joint error. The rounded attention entries remain in $[-1,1]$ because its endpoints belong to this dyadic grid. Every other rounded entry has magnitude at most $p+1$. Its signed integer numerator has magnitude at most $(p+1)2^b$, requiring

$$b + \lceil \log_2(p+2) \rceil + 2 \le C_\beta \log_2 p$$

bits. Use this fixed field length for every dense parameter, including zeros. Here $\log K \le \log p$, and β is fixed. A common header specifying the dimensions and the binary precision costs $O_{d,d',\beta}(\log p)$ additional bits; it is unnecessary if these budget-dependent quantities are known to the decoder. Thus total length is at most $C_\beta p \log_2 p$. Budget inversion gives (57).

Finally, for the finite-code radius in (63), decode each dyadic parameter list to its completed history factor. The resulting finite codebook approximates the whole normalized target ball. For a sufficiently small fixed $c_\beta > 0$, take $p := \lfloor c_\beta B / \log_2 B \rfloor$. For all sufficiently large B, this budget is above the required threshold and its total description fits within B bits. Padding each fixed-length description to B bits makes the codebook cardinality at most 2^B, as required in (63). Since $\log p \asymp \log B$ and $\log\log p \asymp \log\log B$, this proves (67). The architecture and the coefficient, precision, and error bounds are independent of the chosen admissible input modulus; the selected coefficients may depend on the input class and target. □

The upper-bound proof is now complete for an arbitrary admissible modulus ω. For the optimality analysis of Subsection B.5, we now specialize to a Hölder history class. Its multilevel history complexity yields the logarithmic finite-description lower bound. Throughout the remainder of this appendix, we therefore reset $X_n := X_n^{\alpha,L}$, $X_\infty := X_\infty^{\alpha,L}$, $\mathfrak{X} := \mathfrak{X}_{\alpha,L}$, and $\mathcal{M} := \mathcal{M}_{\alpha,L}$.

C.6 Proofs of the history-complexity and capacity bounds

We first identify continuous functions on the history space with valid completed causal families. We then construct many distinguishable Hölder histories, convert them into a finite-description lower bound, and prove the fixed-size infinite-capacity result.

The metric history space. Lemma B.1 shows that Δ_β is a distance inducing the original compact topology on $\mathcal{M}$. Thus the packing number $\mathsf{P}_{\beta;\alpha,L}(r)$ is finite for every $r > 0$, and every Δ_β-Lipschitz factor is continuous in the original history topology.

From history factors to completed families. Let $f : \mathcal{M} \to \mathbb{R}^{d'}$ be continuous, and define

$$F_n^f(z)_i := f(\mu_{n,z,i}), \qquad F_\infty^f(x)(t) := f(\mu_{\infty,x,t}). \tag{129}$$

By Proposition D.5, the prefix-law map $\boldsymbol{\mu} : \mathfrak{X} \to \mathcal{M}$ is continuous; hence $f \circ \boldsymbol{\mu}$ is continuous on $\mathfrak{X}$. This gives continuity of every finite map and joint continuity of the boundary evaluation in (x,t). The latter is uniform on the compact set $X_\infty \times [0,1]$, so $F_\infty^f : X_\infty \to C([0,1];\mathbb{R}^{d'})$ is continuous in the uniform norm. Each finite restriction is prefix-causal because its law uses only the observed prefix. Lemma D.4 therefore shows that (129) is a continuously extendable causal family with path extension F_∞^f. If f is one-Lipschitz for Δ_β, its smooth-test seminorm is at most one.

Proof of Proposition B.7. A multilevel family of admissible histories. We place many token levels in each of many disjoint time cells. Choose nonnegative functions $\psi, \varrho \in C_c^\infty((0,1))$ such that $\psi = 1$ on a neighborhood of $\operatorname{supp}\varrho$ and $\int_0^1 \varrho > 0$. Put $\lambda := \min\{L, 1\}$, and fix constants $0 < c_1 < c_0 \le 1$ with $c_0/c_1 \ge 8$, taking c_0 small enough below. For an integer $q \ge 8$, set

$$h := q^{-1}, \qquad a_h := c_0 \lambda h^\alpha, \qquad \delta_h := c_1 \lambda h, \qquad M_h := 1 + \lfloor a_h/\delta_h \rfloor, \qquad b_j := j\delta_h. \tag{130}$$

Then $M_h \ge 2$, every b_j, $0 \le j < M_h$, lies in $[0, a_h]$, and, for all sufficiently small h,

$$\log M_h \ge c\,[1 + (1-\alpha)\log(1/h)]\,. \tag{131}$$

Let $e_1 := (1,0,\dots,0) \in \mathbb{R}^d$, and let $\psi_k(s) := \psi((s-(k-1)h)/h)$. A level-index vector $\boldsymbol{m} = (m_1,\dots,m_q) \in \{0,\dots,M_h-1\}^q$ selects the amplitude b_{m_k} in time cell $((k-1)h, kh)$. Its

associated path is

$$x_{\boldsymbol{m}}(s) \coloneqq \sum_{k=1}^{q} b_{m_k} \psi_k(s) e_1. \tag{132}$$

The supports lie strictly inside disjoint cells. For distances at most h, the derivative bound gives $\|x_{\boldsymbol{m}}(s) - x_{\boldsymbol{m}}(t)\| \le C a_h h^{-1} |s-t| \le L |s-t|^{\alpha}$ after decreasing c_0; on larger distances the oscillation bound gives the same conclusion. The same choice ensures that the paths lie in Ω. Thus $x_{\boldsymbol{m}} \in X_{\infty}^{\alpha,L}$, and all paths vanish at time one.

We select many index vectors that differ in a fixed fraction of time cells. Write $d_{\mathrm{H}}(\boldsymbol{m}, \boldsymbol{n}) \coloneqq \#\{k : m_k \neq n_k\}$ for their Hamming distance. A standard greedy packing of the discrete cube gives a subset $\Sigma_q \subset \{0, \dots, M_h - 1\}^q$ such that

$$d_{\mathrm{H}}(\boldsymbol{m}, \boldsymbol{n}) \ge q/4, \qquad \log |\Sigma_q| \ge cq \log M_h \quad (\boldsymbol{m} \neq \boldsymbol{n}). \tag{133}$$

Indeed, the Hamming-ball bound at any fixed relative radius below one half is uniform over $M_h \ge 2$. Associate with each selected index vector the terminal history

$$\mu_{\boldsymbol{m}} \coloneqq \int_0^1 \delta_{(x_{\boldsymbol{m}}(s), s)} \, \mathrm{d}s \in \mathcal{M}_{\alpha,L}. \tag{134}$$

All extracted endpoints equal $(0, 1)$.

Pairwise separation by smooth tests. For each pair of selected index vectors, we build a test adapted to the cells in which their amplitude levels differ. Fix a bounded nondecreasing $\vartheta \in C^{\infty}(\mathbb{R})$, constant outside $[-1/4, 1/4]$, with distinct limiting values. For distinct levels u, v, put

$$\eta_{u,v}(x) \coloneqq \operatorname{sgn}(u - v) \vartheta\left(\frac{x - (u+v)/2}{\delta_h} \right). \tag{135}$$

Since $|u - v| \ge \delta_h$, there are constants independent of the levels and of h such that

$$\eta_{u,v}(u) - \eta_{u,v}(v) \ge c, \qquad \left\| \eta_{u,v}^{(j)} \right\|_{\infty} \le C_j \delta_h^{-j}. \tag{136}$$

Write $\varrho_k(s) \coloneqq \varrho((s - (k-1)h)/h)$. For two selected index vectors, define

$$\Phi_{\boldsymbol{m},\boldsymbol{n}}(x, s) \coloneqq c_{\mathrm{test}} \delta_h^{\beta_+} \sum_{k : m_k \neq n_k} \varrho_k(s) \eta_{b_{m_k}, b_{n_k}}(x_1), \tag{137}$$

At each time, at most one summand is active. If $0 < \beta \le 1$, its spatial derivative is $O(\delta_h/\delta_h)$, and its time derivative is $O(\delta_h/h)$. Its uniform norm is $O(\delta_h)$; hence, because $\delta_h \le h$, decreasing c_{test} controls the full $C^1(\mathsf{Q}) = C^{0,1}(\mathsf{Q})$ norm, not only the Lipschitz seminorm.

If $\beta > 1$, put $r_\beta \coloneqq \lceil \beta \rceil - 1$ and $\theta_\beta \coloneqq \beta - r_\beta$. Every mixed derivative with j time and ℓ token derivatives satisfies

$$\left\| \partial_s^j \partial_{x_1}^{\ell} \Phi_{\boldsymbol{m},\boldsymbol{n}} \right\|_{\infty} \le C \delta_h^{\beta} h^{-j} \delta_h^{-\ell} \le C \delta_h^{\beta - j - \ell}, \qquad j + \ell \le r_\beta. \tag{138}$$

Since $\delta_h = c_1 \lambda h$, derivatives of total order r_β are $O(h^{\theta_\beta})$. Rescaling within one cell gives a uniform θ_β-Hölder bound. Moreover, the compact support of ϱ leaves a fixed relative gap: distinct active supports are separated by at least $c_\varrho h$, so the same bound holds across cells. This also covers integer β, for which $\theta_\beta = 1$ and the required endpoint class is $C^{\beta-1,1}$. In either case, after one fixed rescaling,

$$\|\Phi_{\boldsymbol{m},\boldsymbol{n}}\|_{C^{\beta_+}(\mathsf{Q})} \le 1. \tag{139}$$

Because $\psi_k = 1$ on $\operatorname{supp} \varrho_k$, (136) and the Hamming-distance bound give

$$\int_E \Phi_{\boldsymbol{m},\boldsymbol{n}} \, \mathrm{d}(\mu_{\boldsymbol{m}} - \mu_{\boldsymbol{n}}) \ge c \delta_h^{\beta_+} h \, d_{\mathrm{H}}(\boldsymbol{m}, \boldsymbol{n}) \ge c \delta_h^{\beta_+}. \tag{140}$$

Thus $d_{\beta_+}(\mu_{\boldsymbol{m}}, \mu_{\boldsymbol{n}}) \ge c h^{\beta_+}$. Raising this bound to the power β_- gives $\Delta_\beta(\mu_{\boldsymbol{m}}, \mu_{\boldsymbol{n}}) \ge c h^{\beta}$, with constants depending only on the fixed parameters (d, β, L). *From separation to entropy.* Combining this separation with (131)–(133) gives

$$\log \mathsf{P}_{\beta;\alpha,L}(c h^{\beta}) \ge c h^{-1} \left[1 + (1 - \alpha) \log(1/h) \right]. \tag{141}$$

Choosing an integer $q \asymp r^{-1/\beta}$ proves (65).

From histories to target families. Set $Y := \log(B + e^e)$, $Z := 1 + (1-\alpha)\log Y$, and $\varepsilon := c(Z/Y)^\beta$. With $c > 0$ sufficiently small, (65) gives $\log \mathsf{P}_{\beta;\alpha,L}(2\varepsilon) > \log B$ for all large B; also $\varepsilon \le 1$. Choose $N > B$ histories $\mu_1, \dots, \mu_N$ separated by at least 2ε. For each sign vector $\sigma \in \{-1,1\}^N$, prescribe $a_j^\sigma := \varepsilon\sigma_j$ at μ_j. These values are one-Lipschitz for Δ_β, because their pairwise differences are at most 2ε. Their McShane extension is

$$\widetilde{f}_\sigma(\mu) := \min_{1\le j\le N} \left\{ a_j^\sigma + \Delta_\beta(\mu, \mu_j) \right\}. \tag{142}$$

Projecting this scalar function onto $[-\varepsilon, \varepsilon]$ and placing it in the first output coordinate gives a continuous factor $f_\sigma : \mathcal{M} \to \mathbb{R}^{d'}$. It retains the prescribed values and has

$$R_{\beta;\alpha,L}(\mathbf{F}_\sigma) \le 1, \qquad \|f_\sigma\|_\infty \le \varepsilon \le 1, \tag{143}$$

where $\mathbf{F}_\sigma$ is the completed causal family induced by (129). Thus every $\mathbf{F}_\sigma$ belongs to the normalized target ball. Distinct sign vectors give factors separated by 2ε in uniform norm, so a ball of radius strictly below ε contains at most one of these 2^N factors. A codebook with at most $2^B < 2^N$ decoded maps cannot approximate all of them with error below ε. By (63), this proves (66). □

Proof of Proposition B.8. We exhibit an infinite-VC subclass using one scalar head and one hidden ReLU unit. Put $D = d+1$, so the residual width is $r = d+3 = D+2$. In the lifted state $(z, s, 1, w)$, the four blocks are respectively the token, time, homogeneous, and work coordinates. Let e_1^{in} be the first coordinate vector of $\mathbb{R}^d$.

A one-parameter threshold family. For $0 < \vartheta < 1$, choose the single head by

$$Q := \bar{e}_{D+1}^\top, \qquad K := \vartheta \bar{e}_D^\top, \qquad V := \bar{e}_1^\top, \qquad W := \bar{e}_{D+2}, \tag{144}$$

where $(\bar{e}_j)_{j=1}^{D+2}$ is the standard basis of $\mathbb{R}^{D+2}$. Thus Q selects the homogeneous coordinate, K reads time, V reads the first token coordinate, and W writes the result to the work coordinate. On the candidate path $s \mapsto x(s)e_1^{\mathrm{in}}$, the terminal work coordinate is

$$A_\vartheta(x) := \frac{\int_0^1 e^{\vartheta s} x(s)\, \mathrm{d}s}{\int_0^1 e^{\vartheta s}\, \mathrm{d}s}. \tag{145}$$

Let the one-unit readout extract this coordinate, apply ReLU, and write the result into the first output coordinate: explicitly,

$$A_1 := \bar{e}_{D+2}^\top, \qquad b_1 := 0, \qquad A_2 := e_1', \qquad b_2 := 0. \tag{146}$$

Consequently, the first output coordinate is positive exactly when $\Lambda_\vartheta(x) := \int_0^1 e^{\vartheta s} x(s)\, \mathrm{d}s$ is positive. All displayed network weights lie in $[-1, 1]$.

Fixing histories before selecting labels. Fix $N \ge 1$, enumerate the 2^N sign vectors in $\{-1,1\}^N$ as $\sigma^1, \dots, \sigma^{2^N}$, and choose distinct $\vartheta_1, \dots, \vartheta_{2^N} \in (0,1)$. The functionals Λ_{ϑ_ℓ} are linearly independent on $C_c^\infty((0,1))$. Indeed, a relation $\sum_\ell c_\ell \Lambda_{\vartheta_\ell} = 0$ would give

$$\int_0^1 \left(\sum_\ell c_\ell e^{\vartheta_\ell s} \right) x(s)\, \mathrm{d}s = 0 \qquad \text{for every } x \in C_c^\infty((0,1)).$$

The fundamental lemma for test functions makes the parenthesized smooth function vanish on $(0,1)$, and distinct exponentials are linearly independent. The coordinate functionals are therefore independent, so the finite-dimensional range of the linear map

$$x \longmapsto \big(\Lambda_{\vartheta_1}(x), \dots, \Lambda_{\vartheta_{2^N}}(x)\big) \tag{147}$$

has rank 2^N and is onto $\mathbb{R}^{2^N}$. For each $i \in [N]$, choose $\widetilde{x}_i \in C_c^\infty((0,1))$ whose image is the i-th row of the full sign table: $\Lambda_{\vartheta_\ell}(\widetilde{x}_i) = \sigma_i^\ell$ for every ℓ. This choice fixes the same N functions for all 2^N labelings.

Now choose $c_i > 0$ small enough that $x_i := c_i \widetilde{x}_i$ takes values in the unit ball and has α-Hölder seminorm at most L. Positive rescaling changes the magnitudes but preserves every sign. For each of these scaled functions, write

$$x_i^{\mathrm{vec}}(s) := x_i(s) e_1^{\mathrm{in}}, \qquad \mu_{x_i} := \int_0^1 \delta_{(x_i^{\mathrm{vec}}(s), s)} \, \mathrm{d}s. \tag{148}$$

Then $\mu_{x_i} \in \mathcal{M}_{\alpha,L}$ for all i. With these histories fixed, selecting $\vartheta = \vartheta_\ell$ realizes the labeling σ^ℓ, because $\operatorname{sgn} \Lambda_{\vartheta_\ell}(x_i) = \sigma_i^\ell$ for every i. Thus the same N histories are shattered, and, since N was arbitrary, the VC dimension is infinite.

Fixed parameter count. For $H = M = 1$, the dense count (51) is $4(d+3) + (d+3+d'+1) + d' = 5(d+3) + 1 + 2d'$, which proves (69). □

Proof of Corollary B.9. For a prescribed N, use the paths x_i^{vec}, parameters ϑ_ℓ, and labelings σ^ℓ from the preceding proof. They form finite collections, and

$$\Lambda_{\vartheta_\ell}(x_i) = c_i \sigma_i^\ell, \qquad \delta_N := \min_{i \in [N]} c_i > 0.$$

Right-endpoint Riemann sums converge simultaneously for all pairs (i, ℓ). Choose one sufficiently large n so that

$$\left| \frac{1}{n} \sum_{j=1}^{n} e^{\vartheta_\ell j/n} x_i(j/n) - \Lambda_{\vartheta_\ell}(x_i) \right| < \delta_N/2 \qquad \text{for every } i, \ell.$$

The sampled sequences $\mathcal{S}_n x_i^{\mathrm{vec}}$ belong to $X_n^{\alpha,L}$. For parameter ϑ_ℓ, the terminal head numerator has the sign σ_i^ℓ, and its denominator is strictly positive. The one-unit readout in (146) therefore realizes the same threshold labeling as on the path. These sign patterns also ensure that the N sampled sequences are distinct. The heads, readout weights, and parameter count are unchanged, proving the claim for every N. □

D Proof of Theorems 2.4 and 2.6

Both qualitative theorems, including the expanded path statement in Theorem F.1, follow from one argument on the completed evaluation space. We proceed through compactification, target compatibility, prefix-law identification, architectural closure, Laplace separation, one-block realization, and Stone–Weierstrass assembly. The closure argument applies to every fixed finite-depth transformer. It uses step interpolants; Proposition F.2 in Appendix F upgrades the convergence to affine interpolants. Throughout this appendix, write $X_n := X_n^\omega$, $X_\infty := X_\infty^\omega$, and $\mathcal{M} := \mathcal{M}_\omega$ for the fixed common modulus ω.

D.1 Detailed proof reading guide

The argument has four steps, common to the discrete and continuous-time statements. This guide introduces the objects and their roles before the detailed lemmas and proofs below.

1. One compact space of evaluation states. Set

$$\mathfrak{X} := \Big(\bigsqcup_{n \geq 1} \{n\} \times X_n \times [n] \Big) \sqcup \big(\{\infty\} \times X_\infty \times [0,1] \big).$$

A finite state $\xi = (n, z, i)$ records prediction at token i, while $\xi = (\infty, x, t)$ records prediction at path time t. Define its path, time, and resolution coordinates by

$$(P\xi, \tau(\xi), h(\xi)) := \begin{cases} (\mathcal{I}_n z, i/n, 1/n), & \xi = (n, z, i), \\ (x, t, 0), & \xi = (\infty, x, t) \end{cases}$$

and distance

$$d_{\mathfrak{X}}(\xi, \xi') := \|P\xi - P\xi'\|_\infty + |\tau(\xi) - \tau(\xi')| + |h(\xi) - h(\xi')| \, .$$

The coordinate h distinguishes finite resolutions and tends to zero only at the path boundary. The common modulus, the reconstruction property (14), and Arzelà–Ascoli make $\mathfrak{X}$ compact; sampling paths on finer grids makes the finite states dense. The completed target is the joint evaluation

$$\overline{F}^{\star}(n, z, i) \coloneqq F_n^{\star}(z)_i, \qquad \overline{F}^{\star}(\infty, x, t) \coloneqq F_{\infty}^{\star}(x)(t).$$

Definition 2.2 is precisely the condition ensuring that $\overline{F}^{\star}$ is continuous across the finite-to-path boundary. For a path target alone, take $F_n^{\dagger}(z)_i \coloneqq F_{\infty}^{\star}(\mathcal{I}_n z)(i/n)$. Lemmas D.2–D.4 justify this construction, compactness, and target continuity.

2. A complete state variable for the causal history. Let $E \coloneqq \Omega \times [0,1]$, and let $\mathcal{P}(E)$ carry the weak topology (convergence against continuous functions on E). The normalized token–position prefix law of ξ is

$$\mu_\xi \coloneqq \begin{cases} i^{-1} \displaystyle\sum_{j=1}^{i} \delta_{(z_j, j/n)}, & \xi = (n, z, i), \\ t^{-1} \displaystyle\int_0^t \delta_{(x(s), s)} \, \mathrm{d}s, & \xi = (\infty, x, t),\ t > 0, \\ \delta_{(x(0), 0)}, & \xi = (\infty, x, 0). \end{cases}$$

The map $\xi \mapsto \mu_\xi$ is continuous. Its fibers are exactly the causal-information classes: at finite resolution the time marginal has support $\{1/n, \dots, i/n\}$, which determines (n, i), and the token attached to each time recovers $z_{1:i}$. At the boundary the time marginal determines t, and the graph law plus continuity recovers $x|_{[0,t]}$. Hence, with $\mathcal{M} \coloneqq \{\mu_\xi : \xi \in \mathfrak{X}\}$,

$$\mu_\xi = \mu_{\xi'} \iff \xi, \xi' \text{ have the same causal history}, \qquad \overline{F}^{\star}(\xi) = f^{\star}(\mu_\xi)$$

for a unique continuous $f^{\star} : \mathcal{M} \to \mathbb{R}^{d'}$. Here $\mathcal{M}$ is compact, and continuity of the factor follows because a continuous surjection from a compact space onto a Hausdorff space preserves continuous fiberwise factorizations. In particular, resolution is the continuous statistic $h(\xi) = 2 \int s \, \mathrm{d}\mu_\xi - \tau(\xi)$, which vanishes at the path boundary. Proposition D.5 proves the representation, factorization, and continuity of the endpoint and resolution extractors.

3. Separating coordinates computed exactly by attention. For $a \in \mathbb{R}^{d+1}$ with $\|a\| \le 1$, $c \in [0,1]$, and $y = (x, s) \in E$, define

$$\mathcal{L}_{a,c}(\mu) \coloneqq \frac{\int_E e^{c\langle a, y\rangle} \, \langle a, y\rangle \, \mathrm{d}\mu(y)}{\int_E e^{c\langle a, y\rangle} \, \mathrm{d}\mu(y)} = \partial_c \log \int_E e^{c\langle a, y\rangle} \, \mathrm{d}\mu(y).$$

One scalar masked-attention head computes this probe exactly: the homogeneous coordinate supplies query 1, the key and value are $c\langle a, y\rangle$ and $\langle a, y\rangle$, and the output is written to one work coordinate. The same matrices compute the finite softmax average and the boundary temporal integral. Moreover,

$$\log \int_E e^{\langle a, y\rangle} \, \mathrm{d}\mu(y) = \int_0^1 \mathcal{L}_{a,c}(\mu) \, \mathrm{d}c.$$

Thus equality of all probes identifies the Laplace transforms near zero, then every polynomial moment and, because E is compact, the laws. The probes therefore separate points of $\mathcal{M}$ (Lemma D.9).

4. Stone–Weierstrass and realization. Finite sums and products of probes and constants form a point-separating unital algebra on $\mathcal{M}$. Stone–Weierstrass, applied coordinatewise, gives probes $\mathcal{L}_{a_r, c_r}$, $r \in [H]$, and a vector polynomial Π with

$$\sup_{\mu \in \mathcal{M}} \left\| f^{\star}(\mu) - \Pi\big(\mathcal{L}_{a_1, c_1}(\mu), \dots, \mathcal{L}_{a_H, c_H}(\mu)\big) \right\| < \varepsilon/2.$$

One masked-attention block evaluates these probes in parallel, and one shared pointwise ReLU network approximates Π on their compact range to error $\varepsilon/2$. The resulting parameters Θ, independent of n, define a joint finite/boundary output $\overline{T}$ satisfying

$$\sup_{\xi \in \mathfrak{X}} \left\| \overline{T}(\xi) - \overline{F}^{\star}(\xi) \right\| < \varepsilon.$$

The finite and boundary restrictions prove Theorems 2.4 and 2.6, respectively, with the same parameters. Lemma D.10 constructs the parallel heads; Appendix D.8 assembles the polynomial and readout approximation.

D.2 THE JOINT COMPACTIFICATION

The first issue is topological: every finite resolution must fit into one compact space whose only cross-resolution accumulation boundary is the common-modulus path class.

Let $S := \{0\} \cup \{1/n : n \in \mathbb{N}\}$ with the Euclidean metric.

The completed evaluation space is

$$\mathfrak{X} := \left(\bigsqcup_{n \geq 1} \{n\} \times X_n \times [n] \right) \sqcup (\{\infty\} \times X_\infty \times [0,1]) . \tag{149}$$

For $\xi = (n, z, i)$, set $P\xi := \mathcal{I}_n z$, $\tau(\xi) := i/n$, and $h(\xi) := 1/n$; for $\xi = (\infty, x, t)$, set $P\xi := x$, $\tau(\xi) := t$, and $h(\xi) := 0$. Equip $\mathfrak{X}$ with

$$d_{\mathfrak{X}}(\xi, \xi') := \|P\xi - P\xi'\|_\infty + |\tau(\xi) - \tau(\xi')| + |h(\xi) - h(\xi')| . \tag{150}$$

The associated finite-to-boundary convergence criterion is

$$(n_k, z^k, i_k) \to (\infty, x, t) \quad \Longleftrightarrow \quad n_k \to \infty, \quad \mathcal{I}_{n_k} z^k \to x, \quad i_k/n_k \to t. \tag{151}$$

Lemma D.1 (Common-modulus-preserving reconstruction)**.** *For every $n \geq 1$, the reconstruction and sampling operators satisfy*

$$\mathcal{I}_n X_n \subset X_\infty, \qquad \mathcal{S}_n(X_\infty) = X_n, \qquad \mathcal{S}_n \circ \mathcal{I}_n = \mathrm{Id}_{X_n} . \tag{152}$$

Moreover, for every $x \in X_\infty$,

$$\|\mathcal{I}_n \mathcal{S}_n x - x\|_\infty \leq \omega(1/n). \tag{153}$$

Proof. Fix $z = (z_1, \ldots, z_n) \in X_n$, put $h := 1/n$, and introduce knot values $w_0 = w_1 := z_1$ and $w_j := z_j$ for $2 \leq j \leq n$. Convexity of Ω ensures that $\mathcal{I}_n z$ remains in Ω. The discrete common-modulus condition also gives

$$\|w_\ell - w_k\| \leq \omega(h\,|\ell - k|), \qquad 0 \leq k, \ell \leq n; \tag{154}$$

for $k = 0 < \ell$, this follows from $\|z_\ell - z_1\| \leq \omega((\ell - 1)h) \leq \omega(\ell h)$, and all other cases are immediate.

Let $U \sim \mathrm{Unif}[0,1)$ and, for $r \in [0,1]$, define the randomized rounding

$$K_r := \lfloor nr \rfloor + \mathbf{1}_{\{U < nr - \lfloor nr \rfloor\}}.$$

Then K_r is supported on the grid indices bracketing nr, $\mathbb{E}K_r = nr$, and $(\mathcal{I}_n z)(r) = \mathbb{E}w_{K_r}$; this also holds on $[0, h]$ because $w_0 = w_1$. Equivalently, $K_r = \lceil nr - U \rceil$, so $r \mapsto K_r$ is nondecreasing and $s \leq t$ implies $K_s \leq K_t$ almost surely. Fix $0 \leq s \leq t \leq 1$. Since $h(K_t - K_s) \in [0,1]$, (154) and Jensen's inequality for the concave function ω give

$$\begin{aligned} \|(\mathcal{I}_n z)(t) - (\mathcal{I}_n z)(s)\| &\leq \mathbb{E}\,\|w_{K_t} - w_{K_s}\| \\ &\leq \mathbb{E}\,\omega(h(K_t - K_s)) \\ &\leq \omega(h\,\mathbb{E}[K_t - K_s]) = \omega(t - s). \end{aligned}$$

The reverse ordering follows by symmetry. Thus $\mathcal{I}_n z \in X_\infty$, and evaluation at the grid points gives $\mathcal{S}_n \mathcal{I}_n z = z$.

Now take $x \in X_\infty$. Its grid samples obey the all-pairs condition in (11), so $\mathcal{S}_n x \in X_n$. Together with the preceding right-inverse identity, this proves (152). On $[0, h]$, the reconstruction $\mathcal{I}_n \mathcal{S}_n x$ equals $x(h)$ and differs from $x(t)$ by at most $\omega(h)$. On every later grid cell it is a convex combination of the two endpoint samples, each of which is within $\omega(h)$ of $x(t)$. This proves (153). □

For a continuous causal path target $F^\star_\infty$, define its canonical finite restrictions by

$$F^\dagger_n(z)_i := F^\star_\infty(\mathcal{I}_n z)(i/n). \tag{155}$$

Lemma D.2 (Canonical completion of a continuous causal target)**.** *Let $F^\star_\infty : X_\infty \to C([0,1]; \mathbb{R}^{d'})$ be continuous and causal, and define $F^\dagger_n$ by (155). Then every $F^\dagger_n$ is continuous and prefix-causal, and the discrete family $(F^\dagger_n)_{n \geq 1}$ is a continuously extendable causal family with continuous path extension $F^\star_\infty$ in the sense of Definition 2.2.*

Proof. Continuity follows from continuity of $\mathcal{I}_n$, $F^\star_\infty$, and evaluation at the finitely many grid points. If two inputs agree through index i, their affine interpolants agree on $[0, i/n]$; causality of $F^\star_\infty$ therefore gives equality of their i-th outputs.

It remains to prove the uniform extension condition. By Arzelà–Ascoli, X_∞ is compact in the uniform norm. Hence, the image $\mathcal{K} := F^\star_\infty(X_\infty)$ is compact in $C([0,1];\mathbb{R}^{d'})$, hence uniformly equicontinuous. With

$$\omega_{\mathcal{K}}(r) := \sup_{g\in\mathcal{K}} \sup_{|s-t|\le r} \|g(s) - g(t)\|,$$

we have $\omega_{\mathcal{K}}(r) \to 0$ as $r \downarrow 0$. For $z \in X_n$, set $g := F^\star_\infty(\mathcal{I}_n z) \in \mathcal{K}$. By definition, $F^\dagger_n(z) = \mathcal{S}_n g$, and affine reconstruction on each grid cell gives

$$\left\|\mathcal{I}_n F^\dagger_n(z) - F^\star_\infty(\mathcal{I}_n z)\right\|_\infty = \|\mathcal{I}_n \mathcal{S}_n g - g\|_\infty \le \omega_{\mathcal{K}}(1/n).$$

Taking the supremum over X_n proves (15). □

Lemma D.3 (Compact all-resolution state space)**.** *The function $d_{\mathfrak{X}}$ in (150) is a metric, $(\mathfrak{X}, d_{\mathfrak{X}})$ is compact, and the finite states are dense. The finite-to-boundary convergence criterion is exactly (151).*

Proof. Embedding and metric. The path space X_∞ is compact in the uniform norm by Arzelà–Ascoli: its paths are uniformly bounded and share the modulus ω, while the defining conditions are closed under uniform convergence. Define

$$J : \mathfrak{X} \longrightarrow X_\infty \times [0,1] \times S, \qquad J(\xi) := (P\xi, \tau(\xi), h(\xi)). \tag{156}$$

Equip the product with the ℓ^1 metric

$$d_{\mathrm{prod}}\big((x,t,h),(x',t',h')\big) := \|x - x'\|_\infty + |t - t'| + |h - h'|.$$

The map is injective. Indeed, the last coordinate identifies either a finite length n or the boundary. At fixed finite n, the affine interpolant recovers every z_j at j/n, and the normalized evaluation time recovers the index.

Closed image. The image of J is closed. Consider a convergent sequence of image points. If its resolution code converges to $1/n > 0$, that code is isolated in S, so the lengths are eventually equal to n. The finite set of evaluation times is closed. Moreover, X_n is a closed subset of the compact set Ω^n, so it is compact and its continuous image $\mathcal{I}_n X_n$ is compact, hence closed, in X_∞. The limit is therefore again a length-n image point. If the resolution code converges to zero, the limiting path and time already define a boundary state. Thus $J(\mathfrak{X})$ is a closed subset of a compact metric space and is compact. The pullback of this ℓ^1 product metric is precisely (150), proving both positive definiteness and compactness. The same coordinates give (151).

Density of finite states. For density, fix (∞, x, t). Set

$$z^n_j := x(j/n), \qquad i_n := \max\{1, \min\{n, \mathrm{round}(nt)\}\}.$$

Then $z^n \in X_n$, $\|\mathcal{I}_n z^n - x\|_\infty \le \omega(1/n)$, and $|i_n/n - t| \le 1/n$. Hence, $(n, z^n, i_n) \to (\infty, x, t)$. □

D.3 Compatible targets and inherited boundary causality

We next translate continuity on the compactified state space into the uniform output convergence needed by the theorem and show that finite causality survives at the boundary.

For a discrete family and a candidate path extension, define the joint evaluation by

$$\overline{F}^\star(\xi) := \begin{cases} F^\star_n(z)_i, & \xi = (n, z, i), \\ F^\star_\infty(x)(t), & \xi = (\infty, x, t). \end{cases} \tag{157}$$

The required boundary causality is

$$x|_{[0,t]} = x'|_{[0,t]} \quad \Longrightarrow \quad F^\star_\infty(x)(t) = F^\star_\infty(x')(t). \tag{158}$$

Lemma D.4 (Target compatibility and inherited boundary causality)**.** *Assume that every $F_n^\star$ is continuous and that $F_\infty^\star : X_\infty \to C([0,1];\mathbb{R}^{d'})$ is continuous. Then the following are equivalent:*

(a) the joint evaluation map $\overline{F}^\star$ in (157) is continuous;

(b) for every $n_k \to \infty$, $z^k \in X_{n_k}$, and $x \in X_\infty$ with $\mathcal{I}_{n_k} z^k \to x$,

$$\mathcal{I}_{n_k}\big(F_{n_k}^\star(z^k)\big) \longrightarrow F_\infty^\star(x) \quad \text{uniformly on } [0,1]. \tag{159}$$

(c) the uniform extension condition (15) holds.

If the uniform extension condition holds, its continuous path extension is unique. If, in addition, the finite family is prefix-causal, then any of these conditions implies that $F_\infty^\star$ is causal in the sense of (158).

Proof. From joint continuity to uniform path convergence. Suppose first that $\overline{F}^\star$ is continuous. Since $\mathfrak{X}$ is compact, it is uniformly continuous. If $\mathcal{I}_{n_k} z^k \to x$, then uniformly in $i \in [n_k]$,

$$d_{\mathfrak{X}}\big((n_k, z^k, i), (\infty, x, i/n_k)\big) = \big\|\mathcal{I}_{n_k} z^k - x\big\|_\infty + 1/n_k \longrightarrow 0. \tag{160}$$

Consequently,

$$\max_{i\in[n_k]} \big\|F_{n_k}^\star(z^k)_i - F_\infty^\star(x)(i/n_k)\big\| \longrightarrow 0. \tag{161}$$

To make the interpolation step explicit, put $g := F_\infty^\star(x)$ and let ω_g be its modulus of continuity. The affine interpolation operator is nonexpansive for the maximum norm, so

$$\big\|\mathcal{I}_{n_k} F_{n_k}^\star(z^k) - g\big\|_\infty \le \max_{i\in[n_k]} \big\|F_{n_k}^\star(z^k)_i - g(i/n_k)\big\| + \omega_g(1/n_k).$$

Together with (161), this proves (159).

From uniform path convergence to joint continuity. Conversely, consider a convergent sequence in $\mathfrak{X}$. At a finite limit, the resolution code and normalized evaluation time force n and i to be eventually constant, so continuity follows from continuity of $F_n^\star$. A sequence converging to a boundary state may alternate between boundary and finite terms. On the boundary subsequence, convergence follows from continuity into the uniform norm and continuity of evaluation. On the finite subsequence, write $(n_k, z^k, i_k) \to (\infty, x, t)$. Condition (b) and $i_k/n_k \to t$ give

$$F_{n_k}^\star(z^k)_{i_k} = (\mathcal{I}_{n_k} F_{n_k}^\star(z^k))(i_k/n_k) \longrightarrow F_\infty^\star(x)(t).$$

Both subsequences have the same limit, so the full sequence converges and $\overline{F}^\star$ is continuous.

Equivalent extension criterion. It remains to compare (b) and (c). Condition (c) implies (b) because

$$\begin{aligned}&\big\|\mathcal{I}_{n_k} F_{n_k}^\star(z^k) - F_\infty^\star(x)\big\|_\infty \\ &\quad\le \sup_{w\in X_{n_k}} \big\|\mathcal{I}_{n_k} F_{n_k}^\star(w) - F_\infty^\star(\mathcal{I}_{n_k} w)\big\|_\infty + \big\|F_\infty^\star(\mathcal{I}_{n_k} z^k) - F_\infty^\star(x)\big\|_\infty,\end{aligned}$$

and both terms tend to zero. Conversely, if (c) failed, there would be $\varepsilon > 0$ and lengths $n_k \to \infty$ for which the supremum in (15) is at least ε. Choose $z^k \in X_{n_k}$ with discrepancy at least $\varepsilon/2$. Compactness of X_∞ gives, after extraction, $\mathcal{I}_{n_k} z^k \to x$. Condition (b) makes the first path in that discrepancy converge to $F_\infty^\star(x)$, while continuity of $F_\infty^\star$ does the same for the second, a contradiction. Hence (b) and (c) are equivalent.

Uniqueness. Let G_∞ and $\widetilde{G}_\infty$ be two continuous path extensions of the same discrete family. For $x \in X_\infty$, set $z^n := \mathcal{S}_n x$ and $A_n := \mathcal{I}_n F_n^\star(z^n)$. By Lemma D.1, $\mathcal{I}_n z^n \to x$, while the two extension conditions and the triangle inequality give

$$\Big\|G_\infty(\mathcal{I}_n z^n) - \widetilde{G}_\infty(\mathcal{I}_n z^n)\Big\|_\infty \le \|G_\infty(\mathcal{I}_n z^n) - A_n\|_\infty + \Big\|A_n - \widetilde{G}_\infty(\mathcal{I}_n z^n)\Big\|_\infty \longrightarrow 0.$$

Continuity of both extensions yields $G_\infty(x) = \widetilde{G}_\infty(x)$.

Inherited boundary causality. Suppose $x|_{[0,t]} = x'|_{[0,t]}$. If $t > 0$, sample $z_j^n := x(j/n)$ and $(z')_j^n := x'(j/n)$, and set $i_n := \lfloor nt \rfloor$ for all sufficiently large n. The two discrete prefixes agree through i_n, and the corresponding states converge to (∞, x, t) and (∞, x', t). Finite causality and continuity of $\overline{F}^\star$ yield equality of the two boundary values. At $t = 0$, instead use shifted samples $z_j^n := x((j-1)/n)$ and $(z')_j^n := x'((j-1)/n)$, and compare the first outputs. Define $q_n(s) := x((s - 1/n)_+)$ and $q_n'(s) := x'((s - 1/n)_+)$. These shifted paths lie in X_∞ because the time shift is one-Lipschitz and ω is nondecreasing. The displayed sequences are $\mathcal{S}_n q_n$ and $\mathcal{S}_n q_n'$, hence belong to X_n. Lemma D.1 gives

$$\|\mathcal{I}_n z^n - x\|_\infty \le \|\mathcal{I}_n \mathcal{S}_n q_n - q_n\|_\infty + \|q_n - x\|_\infty \le 2\omega(1/n),$$

where $\|q_n - x\|_\infty \le \omega(1/n)$. The same estimate holds for x'. Therefore $(n, z^n, 1) \to (\infty, x, 0)$ and $(n, (z')^n, 1) \to (\infty, x', 0)$. Their first tokens agree because $x(0) = x'(0)$; finite prefix causality and continuity of $\overline{F}^\star$ yield equality of the two boundary values at $t = 0$. □

D.4 Prefix laws identify causal histories

The proof needs a concrete state variable that forgets the future without losing any causal information; token–position prefix laws provide exactly this quotient.

We equip $\mathcal{P}(E)$ with the weak topology. It is compact and Hausdorff because $E = \Omega \times [0, 1]$ is compact metric.

A finite state and a boundary state determine, respectively, the prefix laws

$$\mu_{n,z,i} := \frac{1}{i} \sum_{j=1}^{i} \delta_{(z_j, j/n)} \tag{162}$$

and

$$\mu_{\infty,x,t} := \begin{cases} \dfrac{1}{t} \displaystyle\int_0^t \delta_{(x(s),s)} \, \mathrm{d}s, & t > 0, \\ \delta_{(x(0),0)}, & t = 0. \end{cases} \tag{163}$$

Proposition D.5 (Continuous prefix-law representation)**.** *The map*

$$\boldsymbol{\mu} : \mathfrak{X} \to \mathcal{P}(E), \qquad \boldsymbol{\mu}(\xi) := \mu_\xi, \tag{164}$$

defined in (162)–(163) is continuous. Moreover, $\mu_\xi = \mu_{\xi'}$ if and only if:

(i) $\xi = (n, z, i)$, $\xi' = (n, z', i)$, and $z_{1:i} = z'_{1:i}$; or

(ii) $\xi = (\infty, x, t)$, $\xi' = (\infty, x', t)$, and $x|_{[0,t]} = x'|_{[0,t]}$.

Thus $\mathcal{M} = \boldsymbol{\mu}(\mathfrak{X})$ is compact, and $\boldsymbol{\mu}$ induces a canonical homeomorphism from $\mathfrak{X}/\sim_{\mathrm{pref}}$, equipped with the quotient topology, onto $\mathcal{M}$. The tagged endpoint also descends to a unique continuous map

$$\mathfrak{e} : \mathcal{M} \to E, \qquad \mathfrak{e}(\mu_\xi) = \big(P\xi(\tau(\xi)), \tau(\xi)\big). \tag{165}$$

Writing $\tau(\mu)$ for the time coordinate of $\mathfrak{e}(\mu)$, the resolution code descends to the continuous map $\mathfrak{h} : \mathcal{M} \to S$ given by

$$\mathfrak{h}(\mu) := 2 \int_E s \, \mathrm{d}\mu(x, s) - \tau(\mu), \qquad \mathfrak{h}(\mu_\xi) = h(\xi). \tag{166}$$

For every continuously extendable causal family, completed by its unique path extension, there is a unique continuous

$$f^\star : \mathcal{M} \to \mathbb{R}^{d'} \quad \text{such that} \quad \overline{F}^\star = f^\star \circ \boldsymbol{\mu}. \tag{167}$$

Proof. Continuity on fixed strata. At a finite state, the resolution and evaluation index are locally constant in $\mathfrak{X}$, after which continuity follows directly from continuity of the atoms in (162). On the boundary, let $(x_k, t_k) \to (x, t)$ uniformly and in time. For every $g \in C(E)$, integration against μ_{∞,x_k,t_k} on the subsequence with $t_k = 0$ is evaluation at $(x_k(0), 0)$, which converges to $g(x(0), 0)$.

On the subsequence with $t_k > 0$, uniform continuity of g gives convergence of $t_k^{-1} \int_0^{t_k} g(x_k(s), s)\, \mathrm{d}s$ to the corresponding integral when $t > 0$. If $t = 0$, the entire integration interval shrinks to zero and the paths converge uniformly, so the averages converge to $g(x(0), 0)$. This proves weak continuity on the boundary.

Continuity across resolutions. Now consider finite-to-boundary convergence $(n_k, z^k, i_k) \to (\infty, x, t)$. Let $t_k \coloneqq i_k/n_k$, and define the step path

$$y_k(s) \coloneqq \big(z_j^k, j/n_k\big) \quad \text{for } s \in ((j-1)/n_k, j/n_k], \qquad y_k(0) \coloneqq (z_1^k, 1/n_k). \tag{168}$$

The token component of y_k differs from $\mathcal{I}_{n_k} z^k$ by at most $\omega(1/n_k)$. Hence, $\mathcal{I}_{n_k} z^k \to x$ implies

$$\|y_k - [s \mapsto (x(s), s)]\|_\infty \le \big\|\mathcal{I}_{n_k} z^k - x\big\|_\infty + \omega(1/n_k) + n_k^{-1} \longrightarrow 0. \tag{169}$$

For $t_k > 0$, the empirical law μ_{n_k, z^k, i_k} is exactly the pushforward of $\mathrm{Unif}[0, t_k]$ by y_k. If $t > 0$, (169) and $t_k \to t$ give weak convergence to (163). If $t = 0$, then

$$\sup_{0 \le s \le t_k} \|y_k(s) - (x(0), 0)\| \le \|y_k - [s \mapsto (x(s), s)]\|_\infty + \sup_{0 \le s \le t_k} \big(\|x(s) - x(0)\| + s\big) \longrightarrow 0.$$

Hence, the laws converge to the required Dirac mass. A general sequence approaching a boundary state splits into boundary and finite subsequences; the two arguments above give the same limiting law. Thus $\boldsymbol{\mu}$ is continuous everywhere.

Identification of the fibers. Project a prefix law onto its time coordinate. A finite state gives

$$\frac{1}{i} \sum_{j=1}^{i} \delta_{j/n}. \tag{170}$$

Its minimum support point is $1/n$, which determines n, and its number of atoms then determines i. At each distinct time j/n, equality of the joint laws identifies the unique attached token z_j. Thus equality of two finite laws is equivalent to equality of the tagged finite prefixes.

A boundary state has time marginal $\mathrm{Unif}[0, t]$ when $t > 0$, and δ_0 when $t = 0$. These marginals determine t and cannot equal the finite marginal (170). At $t = 0$, equality of the graph Dirac masses identifies $x(0)$. At $t > 0$, equality of the graph laws implies, after cancelling the common normalization factor $1/t$, that for every continuous $\varphi : [0, 1] \to \mathbb{R}$ and every coordinate ℓ,

$$\int_0^t \varphi(s) x_\ell(s)\, \mathrm{d}s = \int_0^t \varphi(s) x'_\ell(s)\, \mathrm{d}s.$$

Use the explicit continuous extension

$$\varphi(s) \coloneqq \begin{cases} x_\ell(s) - x'_\ell(s), & 0 \le s \le t, \\ x_\ell(t) - x'_\ell(t), & t \le s \le 1. \end{cases}$$

The displayed identity gives $\int_0^t |x_\ell(s) - x'_\ell(s)|^2\, \mathrm{d}s = 0$. Continuity yields equality at every time, so the paths agree on $[0, t]$. This proves the fiber statement.

Quotient, endpoint, and target factorization. Compactness of $\mathcal{M}$ follows because it is the continuous image of compact $\mathfrak{X}$. Let $q : \mathfrak{X} \to \mathfrak{X}/\!\sim_{\mathrm{pref}}$ be the quotient map. The fiber statement gives a unique bijection $\widetilde{\boldsymbol{\mu}}$ satisfying $\boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}} \circ q$. It is continuous by the definition of the quotient topology. Its domain is compact as a quotient of compact $\mathfrak{X}$, while $\mathcal{M}$ is Hausdorff; hence it is a homeomorphism.

The tagged endpoint map

$$\xi \longmapsto \big(P\xi(\tau(\xi)), \tau(\xi)\big)$$

is continuous: this follows from uniform convergence of the path coordinate, convergence of the time coordinate, and continuity of the limiting path. The fiber characterization shows that it is constant on the fibers of $\boldsymbol{\mu}$. Since $\boldsymbol{\mu}$ is a quotient map, it therefore descends uniquely to the continuous map $\mathfrak{e}$ in (165).

For a finite prefix, the mean time is $(i + 1)/(2n)$, while its endpoint time is i/n. Their combination in (166) therefore equals $1/n$. At a boundary prefix the mean time is $t/2$, also at $t = 0$ under

the Dirac convention, so the same expression is zero. Continuity follows from weak continuity of the time moment and continuity of the endpoint extractor. This formula is directly accessible to the architecture: one zero-score scalar head with time as its value writes $\int_E s \, \mathrm{d}\mu$ into an initially zero work coordinate, while the residual coordinates retain the current time. A pointwise affine combination then recovers $1/n$ on finite prefixes and zero at the path boundary, without a separate length channel.

Finite prefix causality and the inherited boundary causality from Lemma D.4 show that $\overline{F}^\star$ is constant on every fiber of $\boldsymbol{\mu}$, so $f^\star$ in (167) is well-defined and unique. The continuous surjection $\boldsymbol{\mu}$ from compact $\mathfrak{X}$ to Hausdorff $\mathcal{M}$ is a quotient map; hence $f^\star$ is continuous. □

D.5 Discrete attention closes under fixed-depth compositions

Architectural compatibility rests on one stability fact: masked sums converge uniformly to temporal attention, and this convergence survives every fixed composition.

For $u = (u_1, \dots, u_n) \in (\mathbb{R}^p)^n$, define its step interpolation

$$(\mathcal{C}_n u)(s) := u_j \quad \text{for } s \in ((j-1)/n, j/n], \qquad (\mathcal{C}_n u)(0) := u_1. \tag{171}$$

Lemma D.6 (Uniform stability of masked attention)**.** *Let $n_k \to \infty$, let $u^k \in (\mathbb{R}^p)^{n_k}$, and suppose*

$$\left\| \mathcal{C}_{n_k} u^k - u \right\|_\infty \longrightarrow 0 \tag{172}$$

for some $u \in C([0,1]; \mathbb{R}^p)$. For every fixed set of masked multi-head parameters θ,

$$\left\| \mathcal{C}_{n_k} \big(\mathrm{MAtt}_\theta(u^k) \big) - \mathrm{MAtt}_\theta^\infty(u) \right\|_\infty \longrightarrow 0. \tag{173}$$

Proof. Exact step-path identity. Extend (211) to bounded Borel paths using the same Lebesgue-integral ratios and endpoint convention, and denote this extension by $\mathrm{MAtt}_\theta^{\infty,\mathrm{B}}$. At the grid point i/n_k, each cell of the prefix has length $1/n_k$. Multiplying a discrete softmax numerator and denominator by $1/n_k$ gives the exact identity

$$\mathrm{MAtt}_\theta(u^k)_i = \mathrm{MAtt}_\theta^{\infty,\mathrm{B}}(\mathcal{C}_{n_k} u^k)(i/n_k). \tag{174}$$

Joint continuity of attention evaluation. Suppose $v_k \to v$ uniformly, where each v_k is a bounded Borel path, v is continuous, and $t_k \to t$. We claim that

$$\mathrm{MAtt}_\theta^{\infty,\mathrm{B}}(v_k)(t_k) \longrightarrow \mathrm{MAtt}_\theta^\infty(v)(t). \tag{175}$$

First, $v_k(t_k) \to v(t)$, and for each head the score and score–value integrands

$$s \longmapsto e^{\langle Q_r v_k(t_k), K_r v_k(s) \rangle}, \qquad s \longmapsto e^{\langle Q_r v_k(t_k), K_r v_k(s) \rangle} V_r v_k(s)$$

converge uniformly on $[0,1]$ to their counterparts built from $v(t)$ and $v(s)$. The paths are uniformly bounded, so for some $C < \infty$ all scores lie in $[-C, C]$. If $t > 0$, then eventually $t_k \geq t/2$, and every denominator is at least $(t/2)e^{-C} > 0$. For either the scalar denominator integrand or the vector numerator integrand, uniform convergence $g_k \to g$ gives

$$\left\| \int_0^{t_k} g_k(s) \, \mathrm{d}s - \int_0^t g(s) \, \mathrm{d}s \right\| \leq \min\{t_k, t\} \, \|g_k - g\|_\infty + |t_k - t| \max\{\|g_k\|_\infty, \|g\|_\infty\} \longrightarrow 0.$$

Thus both integrals, and hence their ratios, converge. If $t = 0$, then on indices with $t_k = 0$, the endpoint convention gives $V_r v_k(0) \to V_r v(0)$. On indices with $t_k > 0$, each normalized head output is a positive weighted average of $V_r v_k(s)$ over $[0, t_k]$, whose distance from $V_r v(0)$ is bounded by

$$\|V_r\| \left(\|v_k - v\|_\infty + \sup_{0 \leq s \leq t_k} \|v(s) - v(0)\| \right) \longrightarrow 0.$$

Including the residual term proves (175) at every t.

Uniform convergence on the grid. Suppose the grid error did not tend to zero. Choose offending indices i_k and pass to a subsequence such that $t_k = i_k/n_k \to t$. With $v_k = \mathcal{C}_{n_k} u^k$, (175) and (174) give

$$\mathrm{MAtt}_\theta(u^k)_{i_k} \longrightarrow \mathrm{MAtt}_\theta^\infty(u)(t).$$

Applying the joint-continuity claim to the constant sequence $v_k = u$ shows that $\mathrm{MAtt}_\theta^\infty(u)$ is continuous. Hence $\mathrm{MAtt}_\theta^\infty(u)(t_k) \to \mathrm{MAtt}_\theta^\infty(u)(t)$, contradicting the offending grid error. Hence that error tends to zero uniformly. Finally, $\mathrm{MAtt}_\theta^\infty(u)$ is uniformly continuous; replacing each time by the right endpoint of its grid cell changes it by $o(1)$. This upgrades the grid estimate to the step-path estimate (173). □

Corollary D.7 (Closure of every fixed transformer)**.** *Let T_Θ be any fixed finite composition of masked-attention blocks, shared tokenwise affine maps, and pointwise ReLU networks, and let $T_{\Theta,\infty}$ use the same parameters with temporal attention. Let $n_k \to \infty$, let $z^k \in X_{n_k}$, and suppose $\mathcal{I}_{n_k} z^k \to x \in X_\infty$. Then*

$$\left\| \mathcal{C}_{n_k} T_\Theta(\phi_{n_k}(z^k)) - T_{\Theta,\infty}(\phi_\infty(x)) \right\|_\infty \longrightarrow 0. \tag{176}$$

Proof. The token step path differs from its affine reconstruction by at most $\omega(1/n_k)$. Therefore the embedded input step paths satisfy

$$\left\| \mathcal{C}_{n_k} \phi_{n_k}(z^k) - \phi_\infty(x) \right\|_\infty \le \left\| \mathcal{I}_{n_k} z^k - x \right\|_\infty + \omega(1/n_k) + n_k^{-1} \longrightarrow 0.$$

Lemma D.6 propagates uniform convergence through every attention block. If $G : \mathbb{R}^p \to \mathbb{R}^q$ underlies either a shared tokenwise affine map or a shared pointwise ReLU network, then $\mathcal{C}_n(G(u_1), \dots, G(u_n)) = G \circ \mathcal{C}_n u$. Moreover, every such fixed G is globally Lipschitz, so

$$\|G \circ v_k - G \circ v\|_\infty \le \mathrm{Lip}(G) \, \|v_k - v\|_\infty .$$

Thus uniform convergence propagates through each pointwise block. Induction over the finite list of layers proves (176). □

Corollary D.8 (A fixed transformer is an all-resolution causal family)**.** *Fix parameters Θ for a finite causal transformer with output dimension m, and define*

$$\mathcal{T}_{\Theta,n}(z)_i := T_\Theta(\phi_n(z))_i, \qquad \mathcal{T}_{\Theta,\infty}(x)(t) := T_{\Theta,\infty}(\phi_\infty(x))(t). \tag{177}$$

Then the discrete family $(\mathcal{T}_{\Theta,n})_{n\ge1}$ is a continuously extendable causal family in the sense of Definition 2.2, with continuous path extension $\mathcal{T}_{\Theta,\infty}$.

Proof. At every finite resolution, continuity follows from the finite composition of continuous operations, and prefix causality follows from the mask by induction over the layers. For temporal attention, the change of variables $s = ta$ writes each head ratio, for $t > 0$, as

$$\frac{\int_0^1 e^{\langle Q_r u(t), K_r u(ta)\rangle} V_r u(ta) \, \mathrm{d}a}{\int_0^1 e^{\langle Q_r u(t), K_r u(ta)\rangle} \, \mathrm{d}a}.$$

At $t = 0$, the same expression equals $V_r u(0)$, exactly the endpoint convention (212). If $u_k \to u$ uniformly, choose B such that $\|u_k\|_\infty, \|u\|_\infty \le B$. The numerator and denominator integrands converge uniformly on $[0,1]^2$, and the denominator of head r is bounded below explicitly by

$$\int_0^1 e^{\langle Q_r u_k(t), K_r u_k(ta)\rangle} \, \mathrm{d}a \ge \exp\big(-\|Q_r\| \, \|K_r\| \, B^2\big) > 0.$$

Thus temporal attention maps continuous paths to continuous paths and is continuous in the uniform norm. Shared tokenwise affine maps and pointwise ReLU networks preserve these properties, so $\mathcal{T}_{\Theta,\infty} : X_\infty \to C([0,1]; \mathbb{R}^m)$ is continuous. The temporal family is causal by the same layerwise induction, because its value at time t uses hidden states only on $[0,t]$.

It remains to check cross-resolution compatibility. If $n_k \to \infty$, $z^k \in X_{n_k}$, and $\mathcal{I}_{n_k} z^k \to x$, Corollary D.7 gives uniform convergence of the step-interpolated outputs to $\mathcal{T}_{\Theta,\infty}(x)$. Hence, whenever $i_k/n_k \to t$,

$$\mathcal{T}_{\Theta,n_k}(z^k)_{i_k} \longrightarrow \mathcal{T}_{\Theta,\infty}(x)(t).$$

Together with the finite and boundary continuity already proved, this is continuity of the joint evaluation map on $\mathfrak{X}$. Lemma D.4 converts this joint continuity into the uniform interpolated-output condition in Definition 2.2, so that definition applies. □

D.6 Log-Laplace probes separate prefix laws

Universality now reduces to finding continuous, attention-computable coordinates that separate the compact prefix-law space.

For $a \in \mathbb{R}^{d+1}$ and $c \in [0,1]$, define

$$\mathcal{L}_{a,c}(\mu) := \frac{\int_E e^{c\langle a,y\rangle}\,\langle a,y\rangle\,\mathrm{d}\mu(y)}{\int_E e^{c\langle a,y\rangle}\,\mathrm{d}\mu(y)}. \tag{178}$$

Lemma D.9 (Log-Laplace probes determine a compactly supported law)**.** *For every $a \in \mathbb{R}^{d+1}$ and $c \in [0,1]$, the function $\mathcal{L}_{a,c}$ in (178) is continuous on $\mathcal{P}(E)$. If $\mu, \nu \in \mathcal{P}(E)$ satisfy*

$$\mathcal{L}_{a,c}(\mu) = \mathcal{L}_{a,c}(\nu) \qquad \text{for every } a \in \mathbb{R}^{d+1} \text{ with } \|a\| \le 1, \quad c \in [0,1], \tag{179}$$

then $\mu = \nu$.

Proof. The numerator and denominator in (178) are weakly continuous because their integrands are continuous on compact E, and the denominator is strictly positive. Hence $\mathcal{L}_{a,c}$ is continuous.

For every $b \in \mathbb{R}^{d+1}$, the fundamental theorem of calculus gives the exact identity

$$\log \int_E e^{\langle b,y\rangle}\,\mathrm{d}\mu(y) = \int_0^1 \mathcal{L}_{b,c}(\mu)\,\mathrm{d}c, \tag{180}$$

because the logarithmic moment-generating function vanishes at $c = 0$.

Equality of the probes with $\|a\| \le 1$ therefore gives

$$M_\mu(b) = M_\nu(b) \qquad \text{for every } b \in \mathbb{R}^{d+1} \text{ with } \|b\| \le 1,$$

where

$$M_\mu(b) := \int_E e^{\langle b,y\rangle}\,\mathrm{d}\mu(y)$$

is the multivariate Laplace transform of μ. Since E is compact, differentiation under the integral is justified to every order. The transforms agree on a neighborhood of the origin, so for every multi-index $\alpha \in \mathbb{N}_0^{d+1}$,

$$\partial^\alpha M_\mu(0) = \int_E y^\alpha\,\mathrm{d}\mu(y) = \int_E y^\alpha\,\mathrm{d}\nu(y) = \partial^\alpha M_\nu(0).$$

Hence μ and ν agree on the integrals of all coordinate polynomials. The restrictions of coordinate polynomials to E form a unital point-separating algebra and are therefore uniformly dense in $C(E)$ by the real Stone–Weierstrass theorem. Consequently,

$$\int_E g\,\mathrm{d}\mu = \int_E g\,\mathrm{d}\nu \qquad \text{for every } g \in C(E).$$

It follows that $\mu = \nu$.

□

D.7 Parallel realization by one masked attention layer

Separation becomes useful only if all selected coordinates can be evaluated by one shared shallow block at every resolution.

Lemma D.10 (Exact parallel probes with ReLU readout)**.** *Fix an integer $H \ge 1$, probe parameters $(a_r, c_r)_{r=1}^H$, and a continuous map $\Pi : E \times \mathbb{R}^H \to \mathbb{R}^{d'}$. For every $\delta > 0$, there are an integer $M \ge 1$ and parameters $\Theta = (\theta, \eta)$, independent of n, of the exact shallow form (7)–(9), such that*

$$\sup_{n\ge 1}\ \sup_{z\in X_n}\ \max_{i\in[n]} \left\| T_\Theta(\phi_n(z))_i - \Pi\big((z_i, i/n), \mathcal{L}_{a_1,c_1}(\mu_{n,z,i}), \dots, \mathcal{L}_{a_H,c_H}(\mu_{n,z,i})\big)\right\| < \delta, \tag{181}$$

$$\sup_{x\in X_\infty}\ \sup_{t\in[0,1]} \left\| T_{\Theta,\infty}(\phi_\infty(x))(t) - \Pi\big((x(t), t), \mathcal{L}_{a_1,c_1}(\mu_{\infty,x,t}), \dots, \mathcal{L}_{a_H,c_H}(\mu_{\infty,x,t})\big)\right\| < \delta. \tag{182}$$

If $\|a_r\| \le 1$ for every r, the attention matrices can moreover be chosen with the Euclidean operator-norm bounds

$$\|Q_r\|_{\mathrm{op}}, \quad \|K_r\|_{\mathrm{op}}, \quad \|V_r\|_{\mathrm{op}}, \quad \|W_r\|_{\mathrm{op}} \le 1. \tag{183}$$

Proof. Probe heads. Write $y = (x, s) \in E$, use $p_H = d + H + 2$, and identify the first $d + 1$ coordinates of $\mathbb{R}^{p_H}$ with y. The fixed tokenwise lift $\mathcal{E}_H$ in (7) appends exactly one homogeneous coordinate and H zero work coordinates. For head r, use scalar query, key, and value dimensions and choose

$$\begin{aligned} Q_r &:= (0_{d+1}, 1, 0_H), \quad & K_r &:= (c_r a_r^\top, 0, 0_H), \\ V_r &:= (a_r^\top, 0, 0_H), & W_r &:= e_{d+2+r}. \end{aligned} \tag{184}$$

Thus $Q_r, K_r, V_r \in \mathbb{R}^{1\times p_H}$, $W_r \in \mathbb{R}^{p_H\times 1}$, and

$$Q_r(y, 1, 0_H) = 1, \qquad K_r(y, 1, 0_H) = c_r \langle a_r, y\rangle, \qquad V_r(y, 1, 0_H) = \langle a_r, y\rangle .$$

The column W_r writes only into work coordinate r: its content–position and homogeneous rows are zero, so the residual addition leaves $(y, 1)$ unchanged. The parameter rows in (184) also give $\|Q_r\|_{\mathrm{op}} = \|W_r\|_{\mathrm{op}} = 1$, $\|K_r\|_{\mathrm{op}} = c_r \|a_r\|$, and $\|V_r\|_{\mathrm{op}} = \|a_r\|$, proving (183) when the directions lie in the unit ball.

At a finite state (n, z, i), the query $Q_r(\phi_n(z)_i, 1, 0_H)$ is exactly one. The r-th work residual begins at zero, so the output of head r in that coordinate is exactly

$$\frac{\sum_{j\le i} e^{c_r\langle a_r,(z_j,j/n)\rangle}\,\langle a_r, (z_j, j/n)\rangle}{\sum_{j\le i} e^{c_r\langle a_r,(z_j,j/n)\rangle}} = \mathcal{L}_{a_r,c_r}(\mu_{n,z,i}).$$

The heads have disjoint work ranges. Consequently, the residual block output at (n, z, i) is exactly

$$\Big((z_i, i/n), 1, \mathcal{L}_{a_1,c_1}(\mu_{n,z,i}), \dots, \mathcal{L}_{a_H,c_H}(\mu_{n,z,i})\Big).$$

Thus all probes are computed in parallel while $(y, 1)$ is preserved. With the same matrices, temporal attention gives the integral expression $\mathcal{L}_{a_r,c_r}(\mu_{\infty,x,t})$ for $t > 0$. At $t = 0$, the endpoint rule gives $\langle a_r, (x(0), 0)\rangle$, which is exactly the probe evaluated at the Dirac law in (163).

Shared readout. The joint vector consisting of the current lifted token and all probe values ranges over the compact set

$$K := \big\{\big(\mathfrak{e}(\mu), \mathcal{L}_{a_1,c_1}(\mu), \dots, \mathcal{L}_{a_H,c_H}(\mu)\big) : \mu \in \mathcal{M}\big\} \subset E \times \mathbb{R}^H,$$

because $\mathcal{M}$ is compact and every displayed coordinate is continuous. Choose a compact box $B \subset \mathbb{R}^{p_H-1}$ containing K. Coordinatewise Tietze extension gives a continuous map on B agreeing with Π on K. Apply the classical one-hidden-layer ReLU universal approximation theorem (Leshno et al., 1993) to that extension on B. It gives a finite hidden width M; after inserting a zero column for the homogeneous input coordinate, its matrices have dimensions

$$A_1 \in \mathbb{R}^{M\times p_H}, \quad b_1 \in \mathbb{R}^M, \qquad A_2 \in \mathbb{R}^{d'\times M}, \quad b_2 \in \mathbb{R}^{d'}$$

and $\mathrm{MLP}_\eta(u) = A_2\rho(A_1u+b_1)+b_2$ uniformly approximates $\Pi(y, v)$ to Euclidean error δ whenever $u = (y, 1, v)$ and $(y, v) \in K$. The biases b_1, b_2 are already part of (5). Pointwise application after the attention block proves both estimates. This qualitative invocation gives no target-independent or δ-independent bound on M. □

Lemma D.11 (Sequential width–depth realization of probe polynomials)**.** *Fix probes $(a_r, c_r)_{r=1}^H$ and a vector polynomial $\Pi : \mathbb{R}^H \to \mathbb{R}^{d'}$. For every $\delta > 0$, there is a finite-depth causal transformer S_δ, obtained by alternating masked-attention blocks and shared pointwise one-hidden-layer ReLU maps between a shared tokenwise affine initialization and a shared tokenwise affine readout, such that its intermediate residual width is*

$$p_{\mathrm{seq}} := d + 1 + 3d', \tag{185}$$

every attention block has at most d' scalar heads, and

$$\sup_{n\ge1}\sup_{z\in X_n}\max_{i\in[n]} \big\|S_\delta(\phi_n(z))_i - \Pi\big(\mathcal{L}_{a_1,c_1}(\mu_{n,z,i}), \dots, \mathcal{L}_{a_H,c_H}(\mu_{n,z,i})\big)\big\| < \delta, \tag{186}$$

$$\sup_{x\in X_\infty}\sup_{t\in[0,1]} \big\|S_{\delta,\infty}(\phi_\infty(x))(t) - \Pi\big(\mathcal{L}_{a_1,c_1}(\mu_{\infty,x,t}), \dots, \mathcal{L}_{a_H,c_H}(\mu_{\infty,x,t})\big)\big\| < \delta. \tag{187}$$

The block count depends on the polynomial expansion, while the pointwise hidden widths and parameters may depend on the fixed probes, Π, and δ. Neither the width in (185) nor the per-block head count depends on sequence length. If all directions satisfy $\|a_r\| \le 1$, every scalar head can also satisfy the operator-norm bounds in (183).

Proof. Registers. Write each output coordinate as a finite sum of monomials,

$$\Pi_\ell(v) = b_\ell + \sum_{q=1}^{N_\ell} b_{\ell q} \prod_{s=1}^{m_{\ell q}} v_{r(\ell,q,s)}, \qquad \ell \in [d'], \tag{188}$$

where constant monomials have been absorbed into b_ℓ. Append to the $d+1$ token–position coordinates three d'-dimensional registers (q, p, s), initialized tokenwise as

$$(y, q, p, s) \coloneqq (y, \mathbf{1}_{d'}, \mathbf{1}_{d'}, 0_{d'}). \tag{189}$$

They hold, respectively, the current probes, running products, and completed sums.

Execution schedule. Schedule the finitely many factors in (188), processing the output coordinates in parallel and leaving a coordinate idle when it has no factor at the current stage. For every active output coordinate ℓ, one scalar head uses the ℓ-th probe register as its query, the unchanged token–position coordinates for its key and value, and writes back only to that probe register. If the scheduled factor is $v_r = \mathcal{L}_{a_r,c_r}$, choose rows so that

$$Qu = q_\ell = 1, \qquad Ku' = c_r \langle a_r, y' \rangle, \qquad Vu' = \langle a_r, y' \rangle,$$

and choose W to be the probe-register basis vector for coordinate ℓ. This is the calculation in (184). After the residual addition, that register contains $1 + v_r$. There is at most one active head for each output coordinate, hence at most d' heads in the block. The query row selects the probe register and the output column writes to that register, both with operator norm one. The key and value rows have norms $c_r \|a_r\|$ and $\|a_r\|$. Thus unit-ball probe directions give the asserted per-head bounds here too.

The following pointwise map resets q_ℓ to one and performs the ideal update

$$p_\ell^+ \coloneqq p_\ell(q_\ell - 1)$$

when the monomial continues. At its final factor, it instead adds $b_{\ell q} p_\ell (q_\ell - 1)$ to s_ℓ and resets p_ℓ to one. Idle coordinates are unchanged. These updates are polynomial maps on the finite-dimensional register state. On any compact box containing the exact finite-stage trajectories, a one-hidden-layer ReLU map approximates their nonlinear products uniformly. The identity on the y-coordinates, constant resets, and other affine coordinate updates can simultaneously be represented exactly using ReLU pairs. A final shared tokenwise affine readout returns $(b_\ell + s_\ell)_{\ell=1}^{d'}$.

Uniform error control. Only finitely many stages occur. All exact intermediate states range over a compact set because the probe vector ranges over the compact image of $\mathcal{M}$. The pointwise networks preserve y and reset q exactly, and every later key and value ignores the approximate (p, s) registers. Consequently all later probes remain exact; only the finite sequence of register multiplications and additions propagates approximation error.

Enumerate the finitely many pointwise update stages by $k = 1, \dots, N$. The exact pre- and post-attention states at each stage have compact ranges, so choose compact neighborhoods with positive buffer around all of them. Each ideal polynomial update G_k is Lipschitz on the relevant neighborhood, say with constant L_k. If its pointwise ReLU realization has uniform error η_k and the register error after stage k is e_k, then, as long as the approximate trajectory stays in these neighborhoods,

$$e_k \le L_k e_{k-1} + \eta_k, \qquad e_0 = 0.$$

Iterating gives the explicit bound

$$e_N \le \sum_{k=1}^{N} \eta_k \prod_{j=k+1}^{N} L_j,$$

with an empty product equal to one. The positive buffers and this finite sum allow the η_k to be chosen small enough both to keep every approximate state in its prescribed neighborhood and to make $e_N < \delta$, uniformly over $\mathfrak{X}$. The exact temporal head calculation is identical, including at $t = 0$, so the same construction and the same error propagation prove both (186) and (187). □

D.8 STONE–WEIERSTRASS ASSEMBLY

The final assembly combines separation and polynomial density with exact parallel probes and a ReLU-readout approximation.

Proof of Theorems 2.4 and F.1. Common causal factor. For Theorem 2.4, complete the given discrete family by its unique extension. For Theorem F.1, complete the given continuous target by the canonical finite family from Lemma D.2. In either case, Definition 2.2 and Lemma D.4 make the joint evaluation map $\overline{F}^{\star}$ continuous and give causality on the path boundary. Consequently, Proposition D.5 gives a unique continuous factor

$$f^{\star} : \mathcal{M} \to \mathbb{R}^{d'}$$

such that

$$\overline{F}^{\star}(\xi) = f^{\star}(\mu_{\xi}), \qquad \xi \in \mathfrak{X}. \tag{190}$$

Let $\mathcal{A}$ be the unital real algebra generated on $\mathcal{M}$ by the restrictions of $\mathcal{L}_{a,c}$, with $a \in \mathbb{R}^{d+1}$, $\|a\| \le 1$, and $c \in [0,1]$. Lemma D.9 says that $\mathcal{A}$ separates points of the compact Hausdorff space $\mathcal{M}$. The real Stone–Weierstrass theorem therefore makes $\mathcal{A}$ uniformly dense in $C(\mathcal{M};\mathbb{R})$.

Finite approximation and realization. Apply the scalar density result to every output coordinate with tolerance $\varepsilon/(2\sqrt{d'})$. Only finitely many probes occur in the resulting algebra elements. Hence, there exist $(a_r, c_r)_{r=1}^{H}$, with $a_r \in \mathbb{R}^{d+1}$, $\|a_r\| \le 1$, and $c_r \in [0,1]$, and a vector-valued polynomial $\Pi : \mathbb{R}^{H} \to \mathbb{R}^{d'}$ such that

$$\sup_{\mu \in \mathcal{M}} \left\| f^{\star}(\mu) - \Pi\big(\mathcal{L}_{a_1,c_1}(\mu), \dots, \mathcal{L}_{a_H,c_H}(\mu)\big) \right\| < \frac{\varepsilon}{2}. \tag{191}$$

If no probe occurs because the approximating polynomial is constant, append the dummy probe

$$(a_1, c_1) := (0, 0)$$

and let Π ignore its argument. Thus we may assume $H \ge 1$.

Define

$$\widetilde{\Pi} : E \times \mathbb{R}^{H} \to \mathbb{R}^{d'}, \qquad \widetilde{\Pi}(y, v) := \Pi(v).$$

Lemma D.10, applied to $\widetilde{\Pi}$ with $\delta = \varepsilon/2$, supplies one attention block and one pointwise ReLU network, with a shared parameter list $\Theta = (\theta, \eta)$, that approximate this polynomial simultaneously on every finite state and every boundary state.

By Corollary D.8, the resulting joint transformer evaluation

$$\overline{T}_{\Theta} : \mathfrak{X} \to \mathbb{R}^{d'}$$

is continuous and causal. Proposition D.5 therefore gives a unique continuous history factor

$$t_{\Theta} : \mathcal{M} \to \mathbb{R}^{d'}$$

such that

$$\overline{T}_{\Theta}(\xi) = t_{\Theta}(\mu_{\xi}), \qquad \xi \in \mathfrak{X}. \tag{192}$$

Equivalently,

$$\overline{T}_{\Theta} = t_{\Theta} \circ \boldsymbol{\mu}.$$

More explicitly, the construction in Lemma D.10 chooses the r-th masked-attention head to compute $\mathcal{L}_{a_r,c_r}(\mu)$ exactly and to store it in the r-th work register, while the residual coordinates preserve the endpoint $\mathfrak{e}(\mu)$. Thus the pointwise readout supplied by that lemma satisfies

$$t_{\Theta}(\mu) = \mathrm{MLP}_{\eta}\left(\mathfrak{e}(\mu), 1, \mathcal{L}_{a_1,c_1}(\mu), \dots, \mathcal{L}_{a_H,c_H}(\mu)\right), \qquad \mu \in \mathcal{M}. \tag{193}$$

This is precisely the history factor induced by the shared architecture

$$T_{\Theta} = \mathrm{MLP}_{\eta} \circ \mathrm{MAtt}_{\theta} \circ \mathcal{E}_{H}.$$

The map $(y, v) \mapsto \mathrm{MLP}_{\eta}(y, 1, v)$ approximates $\widetilde{\Pi}(y, v) = \Pi(v)$; hence the residual coordinates retain the endpoint variable, although the target polynomial itself ignores it.

Combining (181)–(182) with (193) gives

$$\sup_{\mu\in\mathcal{M}} \left\| t_\Theta(\mu) - \Pi\big(\mathcal{L}_{a_1,c_1}(\mu),\dots,\mathcal{L}_{a_H,c_H}(\mu)\big)\right\| < \frac{\varepsilon}{2}. \tag{194}$$

Consequently, the triangle inequality and (191) yield

$$\sup_{\mu\in\mathcal{M}} \|f^\star(\mu) - t_\Theta(\mu)\| < \varepsilon.$$

Using (190) and (192), we obtain

$$\begin{aligned}\sup_{\xi\in\mathfrak{X}} \left\|\overline{F}^\star(\xi) - \overline{T}_\Theta(\xi)\right\| &= \sup_{\xi\in\mathfrak{X}} \|f^\star(\mu_\xi) - t_\Theta(\mu_\xi)\| \\ &= \sup_{\mu\in\mathcal{M}} \|f^\star(\mu) - t_\Theta(\mu)\| < \varepsilon.\end{aligned}$$

This is the completed estimate.

In the first case, restricting this estimate to the finite states gives (16). In the second case, restricting it to the boundary states gives (215). These are precisely the two theorem claims. Moreover, when the continuous target is the extension of the discrete family in the first theorem, the single construction above proves both estimates with the same parameter list Θ. □

Remark D.12 (Bounded heads suffice for qualitative universality). The separating directions in the proof lie in the Euclidean unit ball. Lemma D.10 therefore gives both qualitative theorems with all four matrices of every scalar attention head satisfying (183). The same bounds hold for the serialized realization in Lemma D.11. In both realizations, the query is exactly one and the key depends only on $y \in \Omega \times [0,1]$, so every score lies in $[-\sqrt{2},\sqrt{2}]$. Each attention weight on an i-token prefix is therefore at most $e^{2\sqrt{2}}/i$: universality does not require increasingly concentrated attention. The shallow head count, the serialized depth, and all pointwise hidden widths and parameter magnitudes remain unrestricted. These are per-head bounds, not a bounded-weight statement for the whole network.

D.9 The exact all-resolution approximation class

We now prove the converse to Theorem 2.4, characterizing exactly the families approximable uniformly across all lengths. Every fixed transformer has a continuous completed realization, and this property is preserved by uniform approximation over all finite lengths. Continuous extendability is therefore necessary as well as sufficient.

Corollary D.13 (Uniform closure characterization)**.** *Fix an admissible common modulus ω and the dimensions d, d'. For a discrete family $\mathbf{F} = (F_n)_{n\ge1}$, with $F_n : X_n^\omega \to (\mathbb{R}^{d'})^n$, define*

$$\|\mathbf{F}\|_{\mathrm{all}} := \sup_{n\ge1}\sup_{z\in X_n^\omega}\max_{i\in[n]} \|F_n(z)_i\|. \tag{195}$$

In the space of families with finite $\|\cdot\|_{\mathrm{all}}$, let T_{sh} denote the families induced by the shallow architecture (9), let $\mathsf{T}_{\mathrm{fin}}$ denote those induced by arbitrary fixed finite compositions of masked attention, shared tokenwise affine maps, and pointwise ReLU networks, and let C_ω denote the continuously extendable causal families of Definition 2.2. Both transformer classes use the canonical positional lift ϕ_n and parameters independent of n. Each approximant has one finite architecture and parameter list fixed across all lengths. Widths and parameters may vary with the requested accuracy, as may depth in $\mathsf{T}_{\mathrm{fin}}$. Then

$$\overline{\mathsf{T}_{\mathrm{sh}}}^{\|\cdot\|_{\mathrm{all}}} = \overline{\mathsf{T}_{\mathrm{fin}}}^{\|\cdot\|_{\mathrm{all}}} = \mathsf{C}_\omega. \tag{196}$$

The first closure is unchanged if every scalar head is required to satisfy (183). Thus increasing depth does not enlarge the uniform approximation closure; this is not a claim about approximation efficiency.

Proof. Sufficiency: shallow approximation of extendable families. By Lemma D.4, every family in C_ω has a continuous evaluation map on compact $\mathfrak{X}$, so its all-resolution norm is finite. Theorem 2.4 places it in the uniform closure of T_{sh}, even with the per-head bounds by Remark D.12. Also, $\mathsf{T}_{\mathrm{sh}} \subset \mathsf{T}_{\mathrm{fin}}$.

Necessity: uniform limits retain a continuous extension. Suppose $\mathbf{T}^{(k)} \in \mathsf{T}_{\mathrm{fin}}$ converges to $\mathbf{F}$ in $\|\cdot\|_{\mathrm{all}}$. Corollary D.8 gives continuous completed evaluation maps $\overline{T}^{(k)} : \mathfrak{X} \to \mathbb{R}^{d'}$. Finite states are dense by Lemma D.3, so continuity gives the same supremum for the difference norm on finite states and on all of $\mathfrak{X}$:

$$\sup_{\xi \in \mathfrak{X}} \left\| \overline{T}^{(k)}(\xi) - \overline{T}^{(\ell)}(\xi) \right\| = \left\| \mathbf{T}^{(k)} - \mathbf{T}^{(\ell)} \right\|_{\mathrm{all}}.$$

Indeed, each boundary state is a limit of finite states, so its output difference is a limit of finite-state differences and cannot exceed their supremum. Controlling all finite lengths therefore also controls the path limits. Thus the completed maps are uniformly Cauchy and converge uniformly to a continuous map $\overline{F} : \mathfrak{X} \to \mathbb{R}^{d'}$, whose finite restriction is the given family $\mathbf{F}$. Define

$$F_\infty(x)(t) := \overline{F}(\infty, x, t).$$

This is a continuous path for each x. Moreover, uniform continuity of $\overline{F}$ on compact $\mathfrak{X}$, together with

$$d_{\mathfrak{X}}\big((\infty, x, t), (\infty, x', t)\big) = \|x - x'\|_\infty,$$

shows that $F_\infty : X_\infty \to C([0,1]; \mathbb{R}^{d'})$ is continuous in the uniform norm: the same continuity estimate holds at every t. At each fixed length, continuity and prefix causality of F_n follow from uniform convergence of the transformer maps. Lemma D.4 now applies to the continuous joint evaluation $\overline{F}$, and gives the uniform extension condition, uniqueness of the extension, and boundary causality. Therefore $\mathbf{F} \in \mathsf{C}_\omega$, proving all the asserted equalities. □

For example, when $d' = 1$, the family $F_n(z)_i := (-1)^n$ is continuous and prefix-causal at each fixed length, but lies outside this closure. On the zero input, any fixed transformer's terminal outputs satisfy

$$b_n := T_\Theta\big(\phi_n(0, \dots, 0)\big)_n \longrightarrow a \quad \text{for some } a \in \mathbb{R}$$

under grid refinement. Its uniform error against this family is therefore at least $\max\{|a-1|, |a+1|\} \geq 1$, by considering even and odd lengths. The zero transformer attains error 1. The obstruction is cross-resolution discontinuity, not an inability to distinguish a particular finite pair of lengths.

E CONTROLLED ODES: CONTINUOUS EXTENSION AND TARGET REGULARITY

This appendix proves Examples 2.3 and 2.8. A controlled ODE turns an input trajectory into a state trajectory; reconstructing the input tokens and sampling the exact solution yields a discrete causal family at every resolution. The key estimates are uniform over the input class and include the initial interpolation cell. We then identify a right-endpoint discretization of additive dynamics that is β-smooth whenever the additive vector field is C^β.

E.1 CONSTRUCTION AND UNIFORM COMPATIBILITY

We first specify the dynamics and verify all parts of Definition 2.2, without assuming differentiability of the input paths or Lipschitz dependence on the control.

Dynamics driven by a reconstructed control. Fix an admissible common modulus ω, the classes $X_n := X_n^\omega$, $X_\infty := X_\infty^\omega$, and an initial state $y_0 \in \mathbb{R}^{d'}$. Let $b : \mathbb{R}^{d'} \times \Omega \to \mathbb{R}^{d'}$ be continuous and satisfy, for some $L_b \geq 0$,

$$\|b(y,u) - b(y',u)\| \leq L_b \|y - y'\| \qquad (y, y' \in \mathbb{R}^{d'},\ u \in \Omega). \tag{197}$$

For $x \in X_\infty$, define y_x by

$$\dot{y}_x(t) = b(y_x(t), x(t)), \qquad y_x(0) = y_0, \qquad t \in [0,1]. \tag{198}$$

The global state-Lipschitz bound gives a unique $y_x \in C^1([0,1]; \mathbb{R}^{d'})$. Using the reconstruction $\mathcal{I}_n$ and sampling $\mathcal{S}_n$ from Section 2.2, set

$$F_\infty^\star(x) := y_x, \qquad F_n^\star(z) := \mathcal{S}_n y_{\mathcal{I}_n z}, \quad \text{that is,} \quad F_n^\star(z)_i = y_{\mathcal{I}_n z}(i/n). \tag{199}$$

Reconstruction stays in X_∞ by (14). These maps sample exact solutions; no numerical time-stepping error is included in their definition.

Uniform bounds imply extendability. Compactness of the control range bounds the vector field at the origin, and its state-Lipschitz constant then controls every solution and its speed. Define

$$C_0 := \max_{u\in\Omega} \|b(0,u)\|, \qquad R := (\|y_0\| + C_0)e^{L_b}, \qquad M := C_0 + L_b R. \tag{200}$$

Proposition E.1 (Controlled ODE compatibility)**.** *Under (197), the maps (199) form a continuously extendable causal family with continuous path extension $F^\star_\infty$. For every $x \in X_\infty$, $\|y_x\|_\infty \le R$ and $\|\dot y_x\|_\infty \le M$, and for every $n \ge 1$,*

$$\sup_{z\in X_n} \|\mathcal{I}_n(F^\star_n(z)) - F^\star_\infty(\mathcal{I}_n z)\|_\infty \le \frac{M}{n}. \tag{201}$$

The constants R, M are independent of the input, resolution, and admissible common modulus.

Proof. Existence and uniform trajectory bounds. Continuity of b and compactness of Ω give $C_0 < \infty$. The state-Lipschitz bound implies $\|b(y,u)\| \le C_0 + L_b \|y\|$, so local solutions cannot blow up on the fixed horizon. The integral equation and Gronwall's inequality give

$$\|y_x(t)\| \le (\|y_0\| + C_0 t)e^{L_b t} \le R, \qquad \|\dot y_x(t)\| \le C_0 + L_b R = M.$$

Thus every solution exists uniquely on $[0,1]$ and is M-Lipschitz in time.

Continuity in the input control. On the compact set $\overline{B}_{\mathbb{R}^{d'}}(0,R) \times \Omega$, define the uniform control modulus

$$\rho_b(r) := \sup_{\substack{\|y\|\le R,\ u,v\in\Omega \\ \|u-v\|\le r}} \|b(y,u) - b(y,v)\|, \qquad r \ge 0. \tag{202}$$

Uniform continuity gives $\rho_b(r) \to 0$ as $r \downarrow 0$. For $x, x' \in X_\infty$, subtraction of their integral equations yields

$$\|y_x(t) - y_{x'}(t)\| \le L_b \int_0^t \|y_x(s) - y_{x'}(s)\| \, \mathrm{d}s + t\, \rho_b(\|x - x'\|_\infty).$$

Since $t \le 1$, Gronwall gives

$$\|F^\star_\infty(x) - F^\star_\infty(x')\|_\infty \le e^{L_b} \rho_b(\|x - x'\|_\infty). \tag{203}$$

Hence $F^\star_\infty$ is continuous in the uniform norm. For each fixed n, $\mathcal{I}_n$ is continuous (indeed, nonexpansive in the maximum token norm), and sampling is continuous, so $F^\star_n$ is continuous.

Prefix causality. If $z_{1:i} = z'_{1:i}$, then $\mathcal{I}_n z = \mathcal{I}_n z'$ on $[0, i/n]$: the initial cell uses only z_1, and each subsequent cell up to time i/n uses two tokens from this prefix. Uniqueness for the ODE restricted to that interval gives $F^\star_n(z)_i = F^\star_n(z')_i$. Similarly, two continuous controls agreeing on $[0,t]$ have identical state trajectories there, so $F^\star_\infty$ is causal. Although affine reconstruction between grid points uses the next grid value, no token after index i enters the output at time i/n.

Uniform extension at the initial and interior cells. For $z \in X_n$, put $y := y_{\mathcal{I}_n z}$. Then $\mathcal{I}_n(F^\star_n(z)) = \mathcal{I}_n \mathcal{S}_n y$. On an interior grid cell, its value is a convex combination of the two neighboring sampled values of y, each within M/n of $y(t)$. On the initial cell $[0, 1/n]$, it equals $y(1/n)$, and

$$\|y(1/n) - y(t)\| \le M(1/n - t) \le M/n.$$

In particular, the interpolated output at zero need not equal y_0, but its error has the same bound. Taking the supremum over time and input proves (201) and verifies the extension condition in Definition 2.2. □

E.2 Refinement, an explicit instance, and scope

The estimates above also describe what happens when one continuous control is sampled more finely, and clarify the distinction between this qualitative example and the quantitative approximation theorem.

Refining samples of a fixed control. For $x \in X_\infty$, combine (201) and (203) with $\|\mathcal{I}_n\mathcal{S}_n x - x\|_\infty \le \omega(1/n)$. The triangle inequality gives

$$\sup_{x\in X_\infty} \|\mathcal{I}_n F_n^\star(\mathcal{S}_n x) - F_\infty^\star(x)\|_\infty \le \frac{M}{n} + e^{L_b}\rho_b(\omega(1/n)) \longrightarrow 0. \tag{204}$$

If b is also L_u-Lipschitz in the control on the bounded state region, the second term is at most $e^{L_b}L_u\omega(1/n)$. No exact equality of targets at different finite resolutions is required: the reconstruction and the normalized sampling times both depend on n.

The controlled integrator. For $d' = d$, $y_0 = 0$, and $b(y, u) = u$, the path operator is integration. The initial constant cell followed by affine cells gives

$$F_\infty^\star(x)(t) = \int_0^t x(s)\ \mathrm{d}s, \qquad F_n^\star(z)_i = \frac{1}{n}\sum_{j=1}^{i} z_j + \frac{z_1 - z_i}{2n}. \tag{205}$$

Indeed, the integral over $[0, 1/n]$ is z_1/n; each later cell contributes $(z_{j-1} + z_j)/(2n)$. Summing gives the formula, also for $i = 1$, where the correction vanishes. Thus exact integration of the reconstructed control differs from the plain right-endpoint cumulative sum, but their grid outputs differ by at most $1/n$, since $\|z_1 - z_i\| \le 2$. Nonexpansiveness of output interpolation then gives the same continuous path limit for both families. Here $M = 1$, so the compatibility defect is at most $1/n$.

What the example does and does not imply. Theorems 2.4 and 2.6 therefore approximate this family, uniformly at all finite resolutions and on continuous controls, with the same transformer parameters. The bound M/n concerns time reconstruction of the exact output, not approximation as a function of the network budget p. Applying Theorem B.3 additionally requires $R_{\beta;\omega}(\mathbf{F}^\star) < \infty$; continuity of the vector field in its control argument alone does not ensure that condition. The present example establishes the qualitative target hypothesis. The next subsection verifies the quantitative hypothesis for a right-endpoint variant and shows why the discretization matters.

E.3 β-SMOOTH CONTROLLED INTEGRATORS

Additive dynamics provide an explicit bridge from the ODE example to the β-smooth teacher class in Definition 2.7. The key is to choose grid targets that are smooth averages of their token–position laws.

Fix $\beta > 0$ and $g \in C^\beta(\Omega; \mathbb{R}^{d'})$, using the componentwise isotropic restriction norm from $[-1, 1]^d$, with the integer-order convention $C^m = C^{m-1,1}$. Define

$$F_\infty^\star(x)(t) \coloneqq y_0 + \int_0^t g(x(s))\ \mathrm{d}s, \qquad F_n^\star(z)_i \coloneqq y_0 + \frac{1}{n}\sum_{j=1}^{i} g(z_j). \tag{206}$$

These finite outputs solve $\dot{y} = g(\mathcal{C}_n z)$ exactly at grid points, where $\mathcal{C}_n z$ is the right-endpoint step reconstruction from (171). They are right-endpoint quadrature values for the continuous ODE, not its exact solution driven by the affine reconstruction $\mathcal{I}_n z$.

Proposition E.2 (A β-smooth ODE teacher)**.** *The family (206) is continuously extendable for every admissible ω. Its completed history factor is*

$$f^\star(\mu) = y_0 + t(\mu)\int_E g(u)\ \mathrm{d}\mu(u, s), \qquad t(\mu) \coloneqq \mathfrak{e}(\mu)_{d+1}, \tag{207}$$

and

$$R_{\beta;\omega}(\mathbf{F}^\star) \le C_{d,d',\beta}\,\|g\|_{C^\beta(\Omega)}\,. \tag{208}$$

The constant is independent of ω and y_0. The optimal seminorm may vary with ω, since its supremum is taken on $\mathcal{M}_\omega$; the displayed upper bound is uniform over these domains. Centering the outputs by y_0 and scaling places the family in $\mathcal{F}^\beta(\omega)$.

Proof. Put $G \coloneqq \|g\|_\infty$ in the Euclidean output norm, and let ρ_g be its uniform modulus of continuity on Ω. Continuity and causality of each grid map are immediate. The continuous operator is causal

and satisfies $\|F^\star_\infty(x) - F^\star_\infty(x')\|_\infty \le \rho_g(\|x - x'\|_\infty)$. On each cell, $\|\mathcal{C}_n z - \mathcal{I}_n z\| \le \omega(1/n)$, so integration gives

$$\max_{i \le n} \|F^\star_n(z)_i - F^\star_\infty(\mathcal{I}_n z)(i/n)\| \le \rho_g(\omega(1/n)).$$

The continuous output is G-Lipschitz in time. Nonexpansiveness of affine interpolation and its initial-cell estimate therefore imply

$$\sup_{z \in X_n} \|\mathcal{I}_n F^\star_n(z) - F^\star_\infty(\mathcal{I}_n z)\|_\infty \le G/n + \rho_g(\omega(1/n)) \longrightarrow 0. \tag{209}$$

This proves continuous extendability. Formula (207) follows from $t = i/n$ and the normalized prefix law at finite resolution; it also holds for every continuous history, including $t = 0$.

Write $m_\mu := \int g \, \mathrm{d}\mu$, $t := t(\mu)$, and $t' := t(\nu)$. The factorization gives

$$\|f^\star(\mu) - f^\star(\nu)\| \le \|m_\mu - m_\nu\| + G\,|t - t'|\,.$$

For $0 < \beta \le 1$, an optimal W_1 coupling and Jensen's inequality give $\|m_\mu - m_\nu\| \le C_{d,d'}\,\|g\|_{C^\beta}\,W_1(\mu,\nu)^\beta$. Also $|t - t'| \le \min\{1, \delta_e(\mu,\nu)\} \le \delta_e(\mu,\nu)^\beta$. Since $W_1 \asymp d_1$ by Lemma B.1, these bounds give (208) with $\Delta_\beta = d_1^\beta + \delta_e^\beta$. For $\beta > 1$, the functions $(u,s) \mapsto g_j(u)$ are admissible smooth tests after scaling, so $\|m_\mu - m_\nu\| \le C_{d,d',\beta}\,\|g\|_{C^\beta}\,d_\beta(\mu,\nu)$. Together with $|t - t'| \le \delta_e(\mu,\nu)$, this proves the high-order bound. Finally, $\|f^\star - y_0\|_\infty \le G$, and translation does not change the seminorm. Dividing the centered family by $\max\{1, G, R_{\beta;\omega}(\mathbf{F}^\star)\}$ gives the normalized teacher class. □

Why the discretization matters. Smooth dynamics alone do not imply the β-smooth target condition. Even the exact affine-control integrator (205) fails it for every $\beta > 1$ when $d = d' = 1$ and $\omega \not\equiv 0$. To see this, take $n \ge 3$, $h := 1/n$, and $a := \min\{\omega(h)/2, h\} > 0$. Compare the tokens $z := (a, -a, 0, \dots, 0)$ with the zero sequence at $i = n$. Both belong to X_n^ω and have the same endpoint. Their exact affine-control outputs differ by $ah/2$. For $\gamma := \min\{\beta - 1, 1\}$ and any unit C^β test,

$$\int \varphi \, \mathrm{d}(\mu_z - \mu_0) = ah \int_0^1 \left[\partial_u \varphi(ra, h) - \partial_u \varphi(-ra, 2h)\right] \mathrm{d}r, \qquad d_\beta(\mu_z, \mu_0) \le C_\beta a h^{1+\gamma}.$$

The last estimate follows from $a \le h$ and the γ-Hölder bound on $\partial_u \varphi$, using extensions arbitrarily close to the restriction norm. The regularity quotient is therefore at least $c_\beta h^{-\gamma}$, which diverges. In contrast, the right-endpoint outputs in (206) agree for this pair. Thus the two discretizations have the same path limit, but only the right-endpoint family is β-smooth for every $\beta > 1$.

F Continuous-time transformers and path-space universality

This appendix defines the continuous-time realization of masked attention, states its path-space universality theorem, and proves that every fixed finite causal transformer converges to this realization under grid refinement. Throughout this appendix, write $X_n := X_n^\omega$ and $X_\infty := X_\infty^\omega$ for the fixed common modulus ω.

F.1 Temporal attention and the continuous-time realization

Whenever hidden-state interpolants converge uniformly along a grid refinement, each masked softmax average converges to a normalized temporal integral over the observed history. Replacing finite sums by these integrals defines the continuous-time realization with the same learned parameters.

For a continuous path $x \in X_\infty$, define the continuous positional lift

$$\phi_\infty(x)(t) := (x(t), t). \tag{210}$$

For a continuous hidden path $u : [0,1] \to \mathbb{R}^p$, the temporal attention block corresponding to (2) is, for $t > 0$,

$$\mathrm{MAtt}^\infty_\theta(u)(t) := u(t) + \sum_{r=1}^{H} W_r \frac{\int_0^t e^{\langle Q_r u(t), K_r u(s)\rangle} V_r u(s) \, \mathrm{d}s}{\int_0^t e^{\langle Q_r u(t), K_r u(s)\rangle} \, \mathrm{d}s}, \tag{211}$$

with the continuous endpoint value

$$\mathrm{MAtt}_\theta^\infty(u)(0) := u(0) + \sum_{r=1}^{H} W_r V_r u(0). \tag{212}$$

Every denominator in (211) is strictly positive. For continuous u, each normalized head value tends to $V_r u(0)$ as $t \downarrow 0$, so (212) is the unique continuous endpoint value. Replacing each discrete attention block by (211)–(212), while retaining the same fixed lifts, affine maps, and pointwise MLPs, defines the continuous-time transformer $T_{\Theta,\infty}$.

For a continuous path $y : [0,1] \to \mathbb{R}^{d+1}$, the lift and readout act pointwise: $(\mathcal{E}_H y)(t) := (y(t), 1, 0_H)$ and $(\mathrm{MLP}_\eta\, u)(t) := \mathrm{MLP}_\eta(u(t))$. The continuous-time realization of (9) is therefore

$$T_{\Theta,\infty} := \mathrm{MLP}_\eta \circ \mathrm{MAtt}_\theta^\infty \circ \mathcal{E}_H, \tag{213}$$

with the scalar-head and MLP dimensions specified in (8) and (6).

This graph-based causal operator integrates the ordered trajectory $s \mapsto (x(s), s)$ only over $[0,t]$. Unlike content-only unmasked mean-field descriptions, it retains temporal order. Standard unmasked attention averages over a full token law even when normalized position is included; temporal attention instead acts on the restricted and renormalized prefix law at each t.

F.2 Path-space universality

Temporal attention is universal on every compact path class controlled by an admissible common modulus. We restate the path-only guarantee of Theorem 2.6 with all dimensions explicit; no finite-grid target appears in its assumptions or conclusion.

Theorem F.1 (Continuous-time universality for causal paths)**.** *Fix an admissible common modulus ω, with X_∞^ω given by (12). Let $F_\infty^\star : X_\infty \to C([0,1]; \mathbb{R}^{d'})$ be continuous in the uniform norm and causal in the sense that*

$$x|_{[0,t]} = x'|_{[0,t]} \quad \Longrightarrow \quad F_\infty^\star(x)(t) = F_\infty^\star(x')(t). \tag{214}$$

For every $\varepsilon > 0$, there exist integers $H, M \geq 1$ and a parameter list $\Theta = (\theta, \eta)$, with residual width $p_H = d + H + 2$, scalar-head dimensions $Q_r, K_r, V_r \in \mathbb{R}^{1 \times p_H}$, $W_r \in \mathbb{R}^{p_H \times 1}$, and readout dimensions $A_1 \in \mathbb{R}^{M \times p_H}$, $b_1 \in \mathbb{R}^M$, $A_2 \in \mathbb{R}^{d' \times M}$, $b_2 \in \mathbb{R}^{d'}$. Every attention matrix Q_r, K_r, V_r, W_r may be chosen to have operator norm at most one for the Euclidean norms. Using the lift and readout defined above with temporal attention, the transformer $T_{\Theta,\infty}$ is given by (213), and

$$\sup_{x \in X_\infty} \|T_{\Theta,\infty}(\phi_\infty(x)) - F_\infty^\star(x)\|_\infty < \varepsilon. \tag{215}$$

The widths and readout parameter magnitudes may depend on ω, the target, and ε; the per-head norm bounds are uniform in these choices. Normalized time presupposes a fixed terminal horizon. The completed-space proof in Appendix D treats this theorem and Theorem 2.4 simultaneously. When $F_\infty^\star$ extends a discrete family, the discrete and continuous estimates are restrictions of one approximant.

F.3 Uniform closure under grid refinement

The common proof constructs discrete and temporal attention simultaneously. The following stronger architectural statement records that every fixed finite-depth discrete transformer converges uniformly to its temporal realization.

Proposition F.2 (Temporal closure of a fixed causal transformer)**.** *Fix a parameter list Θ, independent of sequence length, defining a finite operator T_Θ by a composition of masked-attention blocks, shared tokenwise affine maps, and pointwise ReLU networks, and let $T_{\Theta,\infty}$ be its temporal realization. If $n_k \to \infty$, $z^k \in X_{n_k}$, and $\mathcal{I}_{n_k} z^k \to x \in X_\infty$ uniformly, then*

$$\mathcal{I}_{n_k} T_\Theta(\phi_{n_k}(z^k)) \longrightarrow T_{\Theta,\infty}(\phi_\infty(x)) \quad \textit{uniformly on } [0,1]. \tag{216}$$

Proof. Step-interpolant convergence. For $u = (u_1, \dots, u_n)$, let $\mathcal{C}_n u$ denote the right-endpoint step interpolation defined in (171). Under the hypotheses of the proposition, Corollary D.7 gives

$$\left\| \mathcal{C}_{n_k} T_\Theta(\phi_{n_k}(z^k)) - T_{\Theta,\infty}(\phi_\infty(x)) \right\|_\infty \longrightarrow 0. \tag{217}$$

Affine-interpolant upgrade. It remains to replace $\mathcal{C}_n$ by the affine interpolant $\mathcal{I}_n$ used in the proposition. More generally, if $u^k \in (\mathbb{R}^m)^{n_k}$, $u \in C([0,1]; \mathbb{R}^m)$, and $\mathcal{C}_{n_k} u^k \to u$ uniformly, then

$$\left\| \mathcal{I}_{n_k} u^k - u \right\|_\infty \le \left\| \mathcal{C}_{n_k} u^k - u \right\|_\infty + \omega_u(1/n_k), \tag{218}$$

where $\omega_u(r) := \sup_{|s-t| \le r} \|u(s) - u(t)\|$. Indeed, on each grid cell the affine interpolant is a convex combination of the two adjacent token values. If $e_k := \left\| \mathcal{C}_{n_k} u^k - u \right\|_\infty$, the left and right token values are within e_k of u at their respective grid points. Their affine combination is therefore within e_k of the corresponding chord of u, and both chord endpoints are within $\omega_u(1/n_k)$ of $u(t)$ at any time t in that cell. The same estimate holds on the initial cell, where $\mathcal{I}_{n_k} u^k$ is constant. The limit path in (217) is continuous, so (218) proves the claim. □

G Proof of the generalization bound

Theorem 2.9 combines bounded-weight approximation, a length-independent parameter cover, and a squared-loss concentration inequality. The statistical observation is an entire sequence: independence is used only between the N training sequences. The architecture-specific step is the cover; the statistical step is a classical Bernstein argument. We then justify raw squared-loss training followed by clipping. Clipping decreases the empirical loss, but need not turn a raw empirical minimizer into a minimizer of the clipped loss.

G.1 History representation and the approximation comparator

Writing the network on the common history space makes one approximation comparator and one parameter cover available at every resolution. Fix the admissible modulus ω and use the compact history space $\mathcal{M}_\omega$, endpoint map $\mathfrak{e}$, and target factor $f^\star$ from Appendix B. Write $E = \Omega \times [0,1]$ and recall that the history at position i is

$$\mu_{n,z,i} = \frac{1}{i} \sum_{j=1}^{i} \delta_{(z_j, j/n)}, \qquad f^\star(\mu_{n,z,i}) = F_n^\star(z)_i.$$

For a fixed width pair (H, M), set $r := d + H + 2$ and $u(y) := (y, 1, 0_H) \in \mathbb{R}^r$. The raw transformer acts on a history as

$$v_\theta(\mu) := u(y) + \sum_{h=1}^{H} W_h \frac{\int_E e^{(Q_h u(y))(K_h u(y'))} V_h u(y') \, \mathrm{d}\mu(y')}{\int_E e^{(Q_h u(y))(K_h u(y'))} \, \mathrm{d}\mu(y')}, \qquad y = \mathfrak{e}(\mu), \tag{219}$$

$$t_\Theta(\mu) := A_2 \rho(A_1 v_\theta(\mu) + b_1) + b_2. \tag{220}$$

At $\mu_{n,z,i}$, this equals $T_\Theta(\phi_n(z))_i$; the normalization by $1/i$ cancels between numerator and denominator. The same formulas give the temporal realization on continuous-path histories. All parameters here have the dimensions specified in Section 2.1.

Let clip be Euclidean projection onto the closed unit ball of $\mathbb{R}^{d'}$. For the raw class $\mathfrak{T}_p$ in Theorem 2.9, introduce its clipped history class

$$\mathcal{G}_p := \{\mathrm{clip} \circ t_\Theta : T_\Theta \in \mathfrak{T}_p\}, \qquad \|g - h\|_{\infty,2} := \sup_{\mu \in \mathcal{M}_\omega} \|g(\mu) - h(\mu)\|. \tag{221}$$

Clipping is a fixed, parameter-free output operation, not an additional learned layer of the shallow architecture. It will be used for statistical analysis and for the final prediction, not in the training objective of the main theorem.

Set

$$q_\beta(p) := \left(\frac{\log \log p}{\log p} \right)^{\beta/(d+2)}. \tag{222}$$

The bounds $R_{\beta;\omega}(\mathbf{F}^\star) \le 1$ and $\|f^\star\|_\infty \le 1$ allow us to apply Corollary B.5. Consequently, for every sufficiently large integer p, there is a deterministic $T_p^\circ \in \mathfrak{T}_p$, with history realization t_p°, such that

$$\left\|t_p^\circ - f^\star\right\|_{\infty,2} \le C_{\mathrm{app}} q_\beta(p). \tag{223}$$

Here and below, the threshold on p and C_{app} depend only on (d, d', β). The corollary controls the error on both discrete and continuous histories and gives exactly the required attention entry bound 1 and readout entry bound $p+1$. Its precision claim is not a restriction on the trained weights: its role here is to supply one admissible approximation comparator. The comparator depends on the target and budget, but not on the training sample. Since clip is nonexpansive and fixes $f^\star(\mu)$, it also gives

$$\inf_{g \in \mathcal{G}_p} \|g - f^\star\|_{\infty,2} \le C_{\mathrm{app}} q_\beta(p).$$

G.2 Uniform stability and covering numbers

The key architectural property is that each attention head is a normalized average. Its sensitivity depends on the parameter bounds, not on the number of tokens or the duration of a continuous prefix.

Lemma G.1 (Stability of normalized exponential averages)**.** *For a probability measure μ and bounded real scores a, b, put $w_a := e^a / \int e^a \, \mathrm{d}\mu$. Then*

$$\int |w_a - w_b| \, \mathrm{d}\mu \le 2 \|a - b\|_\infty . \tag{224}$$

For bounded scalar values v, v', this implies

$$\left| \int v w_a \, \mathrm{d}\mu - \int v' w_b \, \mathrm{d}\mu \right| \le \|v - v'\|_\infty + 2 \|v'\|_\infty \|a - b\|_\infty .$$

Proof. For $s \in [0,1]$, interpolate $a_s := b + s(a-b)$. Differentiating the normalized density gives

$$\partial_s w_{a_s} = w_{a_s} \left(a - b - \int (a-b) w_{a_s} \, \mathrm{d}\mu \right).$$

Boundedness justifies differentiation under the integral. Since $\int w_{a_s} \, \mathrm{d}\mu = 1$, the $L^1(\mu)$ norm of this derivative is at most $2\|a-b\|_\infty$. Integrating over s proves (224). Add and subtract $\int v' w_a \, \mathrm{d}\mu$ for the second assertion. The proof applies to empirical measures of every size and to normalized temporal measures. □

Lemma G.2 (A length-independent parameter cover)**.** *For every $0 < \epsilon \le 1$, the class $\mathcal{G}_p$ admits an ϵ-net in $\|\cdot\|_{\infty,2}$, with centers in $\mathcal{G}_p$, whose cardinality $K_p(\epsilon)$ satisfies*

$$\log K_p(\epsilon) \le 2 \log p + p \log \left(1 + \frac{160 \sqrt{d'} p^9}{\epsilon} \right). \tag{225}$$

For each fixed width pair, the parameter-to-predictor map is $40\sqrt{d'}p^8$-Lipschitz from the parameter maximum norm to $\|\cdot\|_{\infty,2}$, both before and after clipping.

Proof. Fix an admissible pair (H, M). The dense parameter count (51) implies $H, r, M \le p$. Compare two parameter lists in the specified boxes, at coordinatewise distance at most h. Since $\|u(y)\|_\infty \le 1$, each scalar query, key, or value has magnitude at most r and changes by at most rh. A query–key product therefore changes by at most $2r^2 h$. Lemma G.1 bounds the change in a scalar head average by

$$rh + 2r(2r^2 h) = (r + 4r^3)h.$$

The head average itself has magnitude at most r. Each coordinate of W_h has magnitude at most one and changes by at most h. Perturbing both factors, then summing over heads, yields

$$\|v_\theta(\mu)\|_\infty \le 1 + Hr \le 2p^2, \qquad \|v_\theta(\mu) - v_{\theta'}(\mu)\|_\infty \le H(2r + 4r^3)h \le 6p^4 h. \tag{226}$$

These inequalities are uniform in μ.

Write $B := p + 1 \le 2p$ for the readout entry bound. Each hidden preactivation has magnitude at most $rB(2p^2) + B \le 6p^4$. Its change is at most

$$rh(2p^2) + rB(6p^4 h) + h \le 15p^6 h.$$

ReLU is one-Lipschitz and preserves the preceding magnitude bound. Each coordinate of the final affine output consequently changes by at most

$$Mh(6p^4) + MB(15p^6 h) + h \le 37p^8 h.$$

Passing to the Euclidean output norm gives the claimed, slightly enlarged constant $40\sqrt{d'}p^8$. Nonexpansiveness of clip gives the same estimate after clipping.

Cover every parameter interval by intervals of radius $h := \epsilon/(40\sqrt{d'}p^8)$, with centers in that parameter interval. Its length is at most $2(p+1) \le 4p$, so $1 + 160\sqrt{d'}p^9/\epsilon$ centers per coordinate suffice. There are at most p coordinates. Finally, there are at most p^2 admissible integer pairs (H, M); take the union of their covers. This proves (225) without any dependence on ω or context length. □

G.3 Squared loss on independent complete sequences

Conditional centering identifies excess loss with prediction error, while normalization lets us concentrate one bounded loss per sequence. Write $S_k := (n_k, Z^k, Y^k)$ for the training observations and $S := (n, Z, Y)$ for an independent observation with the same law. For a history predictor g, define

$$\ell_g(S) := \frac{1}{n}\sum_{i=1}^{n} \|g(\mu_{n,Z,i}) - Y_i\|^2, \qquad \mathcal{L}(g) := \mathbb{E}\ell_g(S), \qquad \widehat{\mathcal{L}}_N(g) := \frac{1}{N}\sum_{k=1}^{N} \ell_g(S_k). \tag{227}$$

Here the expectation uses a fresh observation from the training law. We use the same notation for raw predictors when needed. Put $\mathcal{R}_{\mathrm{pred}}(g) := \mathcal{L}(g) - \mathcal{L}(f^\star)$, and let $Z_g(S) := \ell_g(S) - \ell_{f^\star}(S)$ and $\widehat{Z}_g := N^{-1}\sum_{k=1}^{N} Z_g(S_k)$. For every fixed bounded causal predictor, conditional centering gives

$$\mathcal{R}_{\mathrm{pred}}(g) = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} \|g(\mu_{n,Z,i}) - f^\star(\mu_{n,Z,i})\|^2. \tag{228}$$

Indeed, expand each squared loss and condition its cross term on $(n, Z_1, \dots, Z_i)$. For random n, the conditional-mean assumption is understood on $\{n \ge i\}$ for each fixed i. Both predictions and the factor $1/n$ are measurable with respect to this information. The assumption $\mathbb{E}[Y_i \mid n, Z_1, \dots, Z_i] = F_n^\star(Z)_i$ makes that cross term zero; conditioning on the full future sequence is not required. Conditioning on n is consistent with the architecture, whose normalized positional features encode the grid resolution. The normalized averages are bounded, so taking expectations of these random-length sums requires no moment condition on n. For the data-dependent predictor $\widehat{g}$, population risk is evaluated with the training sample held fixed. Thus (228) is the risk in (21) for its associated family $\widehat{F}$.

Lemma G.3 (Finite-class fast oracle inequality)**.** *Let $\mathcal{H}$ be a finite class of K measurable causal history predictors with values in the unit ball. With probability at least $1 - \delta$, simultaneously for every $\eta \ge 0$ and every η-approximate empirical minimizer $\widehat{h}$ over $\mathcal{H}$,*

$$\mathcal{R}_{\mathrm{pred}}(\widehat{h}) \le 3 \inf_{h \in \mathcal{H}} \mathcal{R}_{\mathrm{pred}}(h) + 2\eta + 96\frac{\log(2K/\delta)}{N}. \tag{229}$$

Proof. For a unit-ball prediction h, target $f^\star$, and label Y,

$$\left| \|h - Y\|^2 - \|f^\star - Y\|^2 \right| = |\langle h - f^\star, h + f^\star - 2Y\rangle| \le 4\|h - f^\star\|.$$

For a whole sequence, Jensen's inequality for the normalized average gives

$$|Z_h| \le 4, \qquad \mathbb{E}Z_h^2 \le 16\mathcal{R}_{\mathrm{pred}}(h), \qquad |Z_h - \mathbb{E}Z_h| \le 8. \tag{230}$$

The first bound also follows directly because both sequence losses belong to $[0, 4]$. No within-sequence independence is used for the second bound: square the sequence average, apply Jensen, and then use (228).

Bernstein's inequality for the N independent sequences and a union bound over the K predictors give, with $t \coloneqq \log(2K/\delta)/N$, the simultaneous event

$$\left|\widehat{Z}_h - \mathcal{R}_{\mathrm{pred}}(h)\right| \le \sqrt{32\mathcal{R}_{\mathrm{pred}}(h)t} + \frac{16}{3}t \le \frac{1}{2}\mathcal{R}_{\mathrm{pred}}(h) + 22t, \qquad h \in \mathcal{H}. \tag{231}$$

We have enlarged the usual linear Bernstein constant; the last inequality uses $\sqrt{32at} \le a/2 + 16t$. On this event,

$$\mathcal{R}_{\mathrm{pred}}(h) \le 2\widehat{Z}_h + 44t, \qquad \widehat{Z}_h \le \tfrac{3}{2}\mathcal{R}_{\mathrm{pred}}(h) + 22t.$$

For any comparator $h \in \mathcal{H}$, approximate empirical minimization therefore gives

$$\mathcal{R}_{\mathrm{pred}}(\widehat{h}) \le 2\widehat{Z}_h + 2\eta + 44t \le 3\mathcal{R}_{\mathrm{pred}}(h) + 2\eta + 88t.$$

Take the infimum and enlarge 88 to 96. The event did not depend on the chosen minimizer or tolerance. □

Transfer from a net to the real-parameter class. If g, h take values in the unit ball and $\|g - h\|_{\infty,2} \le \epsilon$, then

$$|\ell_g(S) - \ell_h(S)| \le 4\epsilon, \qquad |\mathcal{R}_{\mathrm{pred}}(g) - \mathcal{R}_{\mathrm{pred}}(h)| \le 4\epsilon. \tag{232}$$

Let $\mathcal{H} \subset \mathcal{G}_p$ be the net of Lemma G.2. The lower half of (231), together with (232), yields simultaneously for every $g \in \mathcal{G}_p$, with probability at least $1 - \delta$,

$$\mathcal{R}_{\mathrm{pred}}(g) \le 2\widehat{Z}_g + 12\epsilon + 44\frac{\log(2K_p(\epsilon)/\delta)}{N}. \tag{233}$$

Indeed, choose a net point h within ϵ of g, then use $\mathcal{R}_{\mathrm{pred}}(g) \le \mathcal{R}_{\mathrm{pred}}(h) + 4\epsilon$ and $\widehat{Z}_h \le \widehat{Z}_g + 4\epsilon$.

For comparison, if the empirical objective itself uses clipped outputs, a τ_{opt}-approximate empirical minimizer $\widetilde{g}$ over $\mathcal{G}_p$ has a nearest net point which is a $(\tau_{\mathrm{opt}} + 4\epsilon)$-approximate minimizer over $\mathcal{H}$. Lemma G.3, transfer back to $\widetilde{g}$, and approximation of each comparator by a net point give

$$\mathcal{R}_{\mathrm{pred}}(\widetilde{g}) \le 3 \inf_{g \in \mathcal{G}_p} \mathcal{R}_{\mathrm{pred}}(g) + 2\tau_{\mathrm{opt}} + 24\epsilon + 96\frac{\log(2K_p(\epsilon)/\delta)}{N}. \tag{234}$$

Taking $\epsilon = 1/N$ and using (225) makes the last two terms at most

$$C_{d'}\frac{p\log(C_{d'}pN) + \log(2/\delta)}{N}.$$

This establishes the usual fast oracle inequality for clipped-loss training. The raw-loss theorem requires the additional argument below.

G.4 Raw empirical minimization followed by clipping

The clipped-class concentration estimate and a single bounded raw comparator suffice to analyze the main theorem's ordinary training loss.

Proof of Theorem 2.9. Increase p_0, depending only on (d, d', β), so that the comparator in (223) satisfies

$$\left\|t_p^\circ - f^\star\right\|_{\infty,2} \le C_{\mathrm{app}}q_\beta(p) \le 1, \qquad \left\|t_p^\circ\right\|_{\infty,2} \le 2.$$

It follows from (228) that

$$\mathcal{R}_{\mathrm{pred}}(t_p^\circ) \le C_{\mathrm{app}}^2 q_\beta(p)^2. \tag{235}$$

For this one raw comparator, the squared-loss difference at a position satisfies

$$\left|\left\|t_p^\circ - Y\right\|^2 - \left\|f^\star - Y\right\|^2\right| \le 5\left\|t_p^\circ - f^\star\right\| \le 5.$$

The same Jensen argument as before therefore proves $\mathbb{E}Z_{t_p^\circ}^2 \le 25\mathcal{R}_{\mathrm{pred}}(t_p^\circ)$ and $\left|Z_{t_p^\circ} - \mathbb{E}Z_{t_p^\circ}\right| \le 10$. Bernstein's inequality, with $s \coloneqq \log(4/\delta)/N$, gives an event of probability at least $1 - \delta/2$ on which

$$\begin{aligned}\widehat{Z}_{t_p^\circ} &\le \mathcal{R}_{\mathrm{pred}}(t_p^\circ) + \sqrt{50\mathcal{R}_{\mathrm{pred}}(t_p^\circ)s} + \frac{20}{3}s \\ &\le \tfrac{3}{2}\mathcal{R}_{\mathrm{pred}}(t_p^\circ) + 32s.\end{aligned} \tag{236}$$

Here $\sqrt{50as} \le a/2 + 25s$.

On another event of probability at least $1 - \delta/2$, (233) holds with $\epsilon = 1/N$ and δ there replaced by $\delta/2$. Let $\widehat{t} := t_{\widehat{\Theta}_N}$ be the history realization of the raw approximate empirical minimizer $T_{\widehat{\Theta}_N}$, and put $\widehat{g} := \mathrm{clip} \circ \widehat{t}$. Since all labels belong to the unit ball, projection decreases their squared distance. The raw optimization condition (20) thus implies

$$\widehat{Z}_{\widehat{g}} \le \widehat{Z}_{\widehat{t}} \le \widehat{Z}_{t_p^\circ} + \tau_{\mathrm{opt}}. \tag{237}$$

Combining the two concentration events, whose intersection has probability at least $1 - \delta$, gives

$$\begin{aligned}
\mathcal{R}_{\mathrm{pred}}(\widehat{F}) &= \mathcal{R}_{\mathrm{pred}}(\widehat{g}) \\
&\le 3\mathcal{R}_{\mathrm{pred}}(t_p^\circ) + 2\tau_{\mathrm{opt}} + \frac{12}{N} + 44\frac{\log(4K_p(1/N)/\delta)}{N} \\
&\quad + 64\frac{\log(4/\delta)}{N} \\
&\le C_{d,d',\beta}\left\{\left(\frac{\log\log p}{\log p}\right)^{2\beta/(d+2)} + \frac{p\log(C_{d'}pN) + \log(2/\delta)}{N}\right\} + 2\tau_{\mathrm{opt}}.
\end{aligned}$$

The last step uses (225) and (235). This proves the stated bound after enlarging its constants. In particular, no uniform concentration bound for the potentially large raw losses of all of $\mathfrak{T}_p$ was needed. Only the clipped class and one bounded raw comparator were used.

For $p = \lfloor\sqrt{N}\rfloor$ and sufficiently large N, $q_\beta(p) \le C_{d,\beta}q_\beta(N)$, while

$$\frac{p\log(C_{d'}pN)}{N} = O_{d'}\left(\frac{\log N}{\sqrt{N}}\right) = o\big(q_\beta(N)^2\big).$$

Consequently,

$$\mathcal{R}_{\mathrm{pred}}(\widehat{F})^{1/2} \le C_{d,d',\beta}\left(\frac{\log\log N}{\log N}\right)^{\beta/(d+2)} + C_{d'}\sqrt{\frac{\log(2/\delta)}{N}} + \sqrt{2\tau_{\mathrm{opt}}}, \tag{238}$$

using $\sqrt{a+b+c} \le \sqrt{a} + \sqrt{b} + \sqrt{c}$. This is a conditional test-risk bound on the same event of probability at least $1 - \delta$ over training samples. In particular, the logarithmic root rate is retained whenever $\tau_{\mathrm{opt}} = O(q_\beta(N)^2)$ and confidence is fixed. Every estimate is independent of ω, the law of sequence lengths, and any maximum length. □

Existence and optimization. At fixed p, the class is a finite union of compact parameter boxes, and the raw empirical loss is continuous on each box. Thus an empirical minimum exists. The theorem assumes a measurable choice; measurable positive-tolerance choices can also be obtained by searching a countable dense parameter subset. More concretely, uniform continuity on each box ensures that exhaustive minimization on sufficiently fine finite parameter grids attains any prescribed tolerance $\tau_{\mathrm{opt}} > 0$. This is an existence argument, not an efficient training method. For a practical optimizer, τ_{opt} must bound the gap to the global empirical minimum; a small gradient or a decrease in training loss does not by itself certify this gap. The proof does not give such a guarantee for SGD. To retain the logarithmic statistical rate, it suffices that $\tau_{\mathrm{opt}} = O(q_\beta(N)^2)$ when $p = \lfloor\sqrt{N}\rfloor$, as shown in (238).

G.5 Sampling variants and scope

The proof extends to other independently sampled causal observations, but its loss normalization and same-distribution interpretation remain essential.

Why continuity alone does not give uniform learning rate. This obstruction already occurs for length-one sequences. Given N, choose $2N$ distinct tokens in Ω, give them independent random labels in $\{-1, 1\}$, and sample uniformly from these tokens. Any such label assignment extends to a continuous map $h : \Omega \to [-1, 1]$; the family $F_n^\star(z)_i := h(z_i)e_1$ is continuously extendable and causal, with path extension $h \circ x$ in its first output coordinate. After N noiseless training observations,

the sign at an unseen test token remains independent and uniform. Its conditional expected squared error is at least one, and the probability of an unseen test token is $(1 - 1/(2N))^N \geq 1/2$. Averaging over label assignments shows that, for every learner, some bounded continuous teacher has expected prediction risk at least $1/2$ for this design. Thus no rate tending to zero can hold uniformly over all continuous teachers and sampling laws. Some quantitative class restriction is necessary; the β-smooth target condition is one sufficient choice, not a necessary characterization of learnability.

Independent histories and fully observed paths. The proof is unchanged for iid observations (μ, Y), where $\mu \in \mathcal{M}_\omega$, $\|Y\| \leq 1$, and $\mathbb{E}[Y \mid \mu] = f^\star(\mu)$, using loss $\|g(\mu) - Y\|^2$. Such a design may mix discrete and continuous histories. It also applies to iid, fully observed labeled paths (x, Y), with $x \in X_\infty^\omega$ and loss

$$\ell_g(x, Y) \coloneqq \int_0^1 \|g(\mu_{x,t}) - Y(t)\|^2 \, \mathrm{d}t, \qquad \mu_{x,t} \coloneqq \frac{1}{t} \int_0^t \delta_{(x(s),s)} \, \mathrm{d}s \quad (t > 0),$$

and $\mu_{x,0} \coloneqq \delta_{(x(0),0)}$. Assume joint measurability, $\|Y(t)\| \leq 1$ almost surely for almost every t, and $\mathbb{E}[Y(t) \mid x|_{[0,t]}] = f^\star(\mu_{x,t})$ for almost every t. Conditional centering and Jensen's inequality under the time integral replace their finite-average versions. Sparse observation or approximate quadrature of this loss requires additional error terms.

What the sample size counts. Each complete sequence has total weight one. The bounds need no moment assumption on the random length, because the normalized sequence loss is bounded after clipping. Arbitrary dependence of tokens and labels within an observation is allowed; overlapping windows from one series are not thereby independent observations. Nor does the theorem silently cover the different objective obtained by dividing the dataset's total token loss by its total number of tokens.

Prediction risk versus length extrapolation. For sequence sampling, define the induced history law $\overline{P} \coloneqq \mathbb{E}[n^{-1} \sum_{i=1}^n \delta_{\mu_{n,Z,i}}]$. The conclusion controls $\|\widehat{g} - f^\star\|_{L^2(\overline{P})}$, not uniform error on all histories. Distribution-independent constants do not turn training on short sequences into a guarantee under an arbitrary new long-sequence distribution. If another history law Q satisfies $\mathrm{d}Q/\,\mathrm{d}\overline{P} \leq \kappa$, then one can infer $\|\widehat{g} - f^\star\|_{L^2(Q)}^2 \leq \kappa \mathcal{R}_{\mathrm{pred}}(\widehat{F})$; without coverage assumptions, no such distribution-shift conclusion follows.

The role of the hypotheses. Weight control is essential to the parameter cover proved here. Finite covers at a positive prediction accuracy do not imply finite zero-threshold VC dimension: thresholding is discontinuous, and arbitrarily small classification margins are not controlled by this argument. The fast estimation rate uses bounded squared regression loss and conditional centering; unrestricted cross-entropy is not covered. The target condition is precisely boundedness of the β-smooth target seminorm in Definition B.2, not generic Fréchet smoothness of a path operator. Finally, the theorem is an upper bound only: neither its statistical optimality nor a statistical lower bound follows from the finite-description approximation lower bounds.

H Empirical protocol and full results

This appendix records the sampling, pretrained maps, scale sensitivity, and refinement checks behind Section 3. Raw physical-signal baselines complement the continuous-content and layer-wise text experiments. The main comparison contrasts continuous patch embeddings with BigBird text input embeddings. Together these measurements characterize finite-scale, representation-dependent regularity; they do not verify a common asymptotic Hölder class.

H.1 Estimator interpretation and sensitivity

The maximum in (23) ranges over every starting position at each tested dyadic lag, making it closer to the all-pairs Hölder geometry used in the optimality benchmark than an L^p structure function. The subscript ∞ denotes this maximum, not a continuum-limit estimate. The factor $d^{-1/2}$ reports root-mean-square coordinate displacement and removes a trivial $\sqrt{d}$ factor when coordinates have

comparable scales; it neither makes arbitrary representations metrically comparable nor normalizes them into the theorem's token ball. The raw physical baselines and BigBird input comparison use the endpoint-inclusive grid $t_k := k/(N-1)$ for each observed sequence. The continuous-content comparison preserves the raw window's time axis: a lag of r feature positions at stride s has $h_r := sr/(N_{\mathrm{raw}}-1)$. Changing a fixed time denominator changes the intercept, but matching physical bands is essential when comparing cadences.

Equation (24) is fit separately for every sequence or window. Thus C is a finite-band prefactor in the chosen encoding and RMS convention, not a common theorem-level Hölder constant. A fixed positive rescaling preserves $\widehat{\alpha}_\infty$ and R^2 while rescaling C. The raw-baseline table below reports the mean and sample standard deviation of $\widehat{\alpha}_\infty$ and the mean R^2. As sensitivity checks, we use Theil–Sen slopes and replace the maximum by the 99th percentile of increments. On the BigBird input rows, Theil–Sen and ordinary least squares agree within a few 10^{-3}. The percentile slope is not a uniform Hölder exponent. More generally, a finite-band fit does not certify regularity: the exact RMS-normalized grid Hölder seminorm takes the maximum of $M_\infty(r)/h_r^\alpha$ over every lag, and a low R^2 argues against a single power law even on the fitted band. We do not use regression standard errors from the correlated lag points as uncertainty across recordings.

In Figure 1, every window w is normalized separately: $R_w(r) := M_{\infty,w}(r)/M_{\infty,w}(1)$. The displayed mean and sample SD are computed across these ratios, so the first-lag point is exactly one with zero SD. This compares curve shapes with equal weight per window and leaves each window's fitted slope unchanged.

H.2 Data and sampling

The benchmark uses a fixed sampling rule in every domain. For the six scalar series, we predeclare the dyadic-plus-one lengths reported in Table 4 and retain up to eight time-spread, mutually disjoint complete windows, yielding between three and eight windows per series. The series are Jena air temperature (Max Planck Institute for Biogeochemistry, n.d.), Beijing PM2.5 concentration (Chen, 2015), building appliance consumption (Candanedo, 2017), Metro Interstate traffic volume (Hogue, 2019), ETT transformer-oil temperature, and ETT high-useful-load measurements (Zhou et al., 2021).

ECG comprises centered five-second Lead-I windows from 24 CPSC/PhysioNet records (Perez Alday et al., 2020; Perez Alday et al., 2022; Pollard et al., 2026), converted to millivolts using each record's gain and baseline and antialias-resampled from 500 Hz to 100 Hz. Ground motion comprises 12 earthquake and 12 quiet 30-second, three-component windows from station `CI.PASC.00`, sampled at 100 Hz; waveforms come from the IRIS FDSN dataselect service and event origins from the USGS event service. The text controls are WikiText-103 (Merity et al., 2017), AG News (Zhang et al., 2015), IMDb (Maas et al., 2011), and this manuscript's mathematical LaTeX source. For the first three, we concatenate documents with the tokenizer separator and extract four nonoverlapping $N = 16{,}384$ windows. For the LaTeX control, we join the frozen manuscript snapshot with blank lines and periodically repeat streams shorter than 65,536 tokens. Its four windows are offsets of a synthetic periodic stream, not independent documents; rerunning after manuscript edits changes this control.

The scalar replicates are disjoint windows from individual recorded series, not independent stations or buildings; the two ETT channels share one source. The reported SD is therefore descriptive of the retained collection. Complete window selection can favor well-observed periods, and missing-data fractions in source metadata describe the entire source rather than each selected window. The seismic cache records waveforms, generic event/noise identifiers, channels, and cadence, but lacks the acquisition timestamps, event identifiers, and request manifest needed for exact re-acquisition. Its event/quiet labels and source coverage are consequently less reproducible than the scalar data; the new continuous-content comparison uses only the six scalar series.

H.3 Pretrained representations

Table 2 summarizes the two pretrained input maps. Chronos-Bolt applies a continuous residual ReLU network to standardized measurement patches, as detailed below. BigBird combines 768-dimensional learned word and token-type embeddings with its learned absolute-position table and applies the pretrained input LayerNorm (Zaheer et al., 2020). At $N = 16{,}384$, we linearly interpolate the

learned position vectors over the normalized interval. This is our input-only adaptation, not the RoPE index-rescaling and fine-tuning procedure of Chen et al. (2023); no attention or fine-tuning is applied. The separate depth experiment in Appendix H.8 uses complete models at native context lengths.

The window-dependent normalization used by Bolt is part of the chosen empirical encoder. It does not certify that one fixed representation map has a common theorem-level Hölder prefactor across sequences and resolutions. The saved metadata pins the Bolt and BigBird snapshots and the three streamed dataset revisions by commit hash, and records the Bolt tensor hashes, BigBird LayerNorm epsilon, and software versions.

Table 2: **Pretrained front ends.** Every row uses Equations (23)– (24); d is the increment dimension.

data type	representation	d	embedding map
Scalar series	Chronos-Bolt-tiny	256	Standardized 16-value patch plus 16 observation-mask entries, pretrained residual ReLU MLP; content only, evaluated at stride 16 and diagnostic stride one.
NLP	BigBird-RoBERTa	768	Learned word and token-type embeddings, linearly interpolated learned absolute position, pretrained input LayerNorm.

H.4 The continuous-content comparison

We extract the input patch embedding of Chronos-Bolt-tiny (Ansari et al., 2024), revision `93a8129`, following the pinned official implementation (Amazon, 2024).[1] For each retained scalar window, mean and population standard deviation are computed once and frozen for all subsequent comparisons; a zero standard deviation is replaced by 10^{-5}. Let $p_k \in \mathbb{R}^{16}$ contain 16 consecutive standardized measurements. The observed-value mask is $\mathbf{1} \in \mathbb{R}^{16}$, since all retained windows are complete. With $u_k := [p_k; \mathbf{1}] \in \mathbb{R}^{32}$, the frozen content map is

$$E(p_k) := W_o \,\mathrm{ReLU}(W_h u_k + b_h) + b_o + W_r u_k + b_r, \tag{239}$$

where $W_h \in \mathbb{R}^{1024\times 32}$, $W_o \in \mathbb{R}^{256\times 1024}$, and $W_r \in \mathbb{R}^{256\times 32}$. All six tensors are pretrained and dropout is inactive. No positional vector or attention output is added. Bolt is an encoder–decoder model, so these measurements concern a local content map, not causal processing by its encoder.

Native complete patches are end-aligned at stride 16, after excluding a possible incomplete left patch. Diagnostic dense evaluation applies exactly the same map to every valid 16-sample patch at stride one. Native outputs match the corresponding dense outputs with zero numerical discrepancy in all 29 retained windows. We omit the non-temporal REG token and do not apply the forecasting pipeline's 2,048-sample truncation. Thus "native" specifies the patch cadence and alignment under the shared full-window normalization, not the full forecasting pipeline. Raw endpoints are restricted to the valid dense patch endpoints, excluding the first 15 raw samples; this accounts for small differences from the full-window raw-baseline table below.

All dyadic feature lags with $h_r \le 1/4$ are retained. The primary band is $h_r \le 1/32$, with at least four positive, finite measurements required for a fit. We also report the first four dyads, the coarse band $1/32 \le h_r \le 1/4$, and the matched physical fine band $16 \le sr \le (N_{\mathrm{raw}} - 1)/32$. No band is chosen according to its slope or R^2. The dense and native primary fits in Table 1 include different finest scales; Table 3 removes that difference. Remaining differences arise because the maximum uses more starting positions at dense cadence. Beijing and traffic have only two native fine-band lags, so we leave their estimates unavailable.

Two permutation controls distinguish temporal organization from the patching operation: one seeded permutation of the standardized scalar input before patch extraction, and one of the resulting dense vectors. Input-shuffled mean slopes are 0.005–0.035, whereas output-shuffled means are within 0.003 of zero. These are controls with one permutation per window, not a permutation significance test. Unprojected delay patches $p_k \in \mathbb{R}^{16}$ are also measured: their positive slopes show that the learned map does not create all the observed regularity.

The connection to the Hölder model is an upper-bound preservation property. If x is a scalar L-Hölder-α path, the patch $p(t) := (x(t - \delta_j))_{j=1}^{P}$ uses fixed physical offsets on its valid time interval,

[1] Official source: `chronos-forecasting v2.0.0`, `src/chronos/chronos_bolt.py`.

Table 3: **Matched physical fine band for the same Bolt map.** Mean $\widehat{\alpha}_\infty \pm$ sample SD; the last column counts dyadic lags. Native cadence uses a subset of the dense starting positions.

domain	dense stride 1	native stride 16	lags
Jena weather	$.157 \pm .039$	$.167 \pm .040$	7
Beijing PM2.5	–	–	2
Building appliances	$.020 \pm .022$	$.020 \pm .017$	4
Road traffic	–	–	2
Transformer oil temp.	$.133 \pm .044$	$.175 \pm .041$	4
Transformer load	$.025 \pm .027$	$.043 \pm .055$	4

and E is K-Lipschitz in Euclidean norm, then

$$\frac{\|E(p(t)) - E(p(s))\|_2}{\sqrt{d}} \le K\sqrt{P/d}\, L|t-s|^\alpha. \tag{240}$$

The fixed ReLU map in (239) is Lipschitz. This observation does not imply equality of exponents, and a common constant also requires controlled normalization. Across resolutions, fixed sample-count patches change their physical support; the native frontend is not thereby identified with one fixed delayed-path map.

H.5 Raw physical-signal baselines

Table 4 reports the full-window physical baselines before learned encoding. The scalar curves use each window divided by its mean absolute value, which does not change its fitted slope. ECG uses the resampled waveform in millivolts. Ground motion uses three waveform components, each centered and divided by its window SD plus 10^{-6}. Thus "raw" denotes measurements with continuous preprocessing, not a common cross-domain amplitude convention. Mean maximum-increment slopes range from 0.041 to 0.389; their variability and fit quality qualify any regularity interpretation.

Table 4: **Raw physical-signal worst-increment regressions.** N is waveform length, d the number of measured coordinates, and m the number of windows. Slopes are mean $\pm$ sample SD on $h \le 1/32$; R^2 is the mean fit quality, and $\alpha_{.99}$ the mean 99th-percentile sensitivity slope.

domain	N	d	m	$\widehat{\alpha}_\infty$	R^2	$\alpha_{.99}$
Jena weather	32,769	1	8	$.253 \pm .040$	.89	.421
Beijing PM2.5	1,025	1	6	$.139 \pm .061$	.73	.365
Building appliances	4,097	1	4	$.041 \pm .020$	.61	.101
Road traffic	1,025	1	3	$.172 \pm .019$	.60	.178
Transformer oil temp.	4,097	1	4	$.134 \pm .027$	.85	.291
Transformer load	4,097	1	4	$.129 \pm .025$	.77	.176
Cardiac ECG	500	1	24	$.185 \pm .174$	.53	.261
Ground motion	3,000	3	24	$.389 \pm .281$	.65	.439

A common normalized cutoff spans different physical durations and contains four fine lags for ECG versus eleven for Jena. Percentile slopes can differ substantially from maximum-increment slopes: for example, Jena gives 0.421 versus 0.253. RMS and percentile slopes describe weaker increment statistics and cannot replace the uniform maximum in the theory.

H.6 Calibration and direct refinement checks

We calibrate the same fine-band estimator on 12 independent realizations per synthetic scenario, each of length $4,097$: a smooth sinusoid with random phase; fractional Brownian motion (fBm) with $H = 0.25, 0.50, 0.75$, generated by circulant embedding of fractional Gaussian noise; independent Gaussian noise; and a sinusoid with one height-two jump at a random location in $[0.3, 0.7]$. The raw and delay-patch baselines use the same standardization and valid endpoints as Bolt. For fBm, H is the supremal pathwise exponent: paths are Hölder for every $\alpha < H$, not generally at H itself.

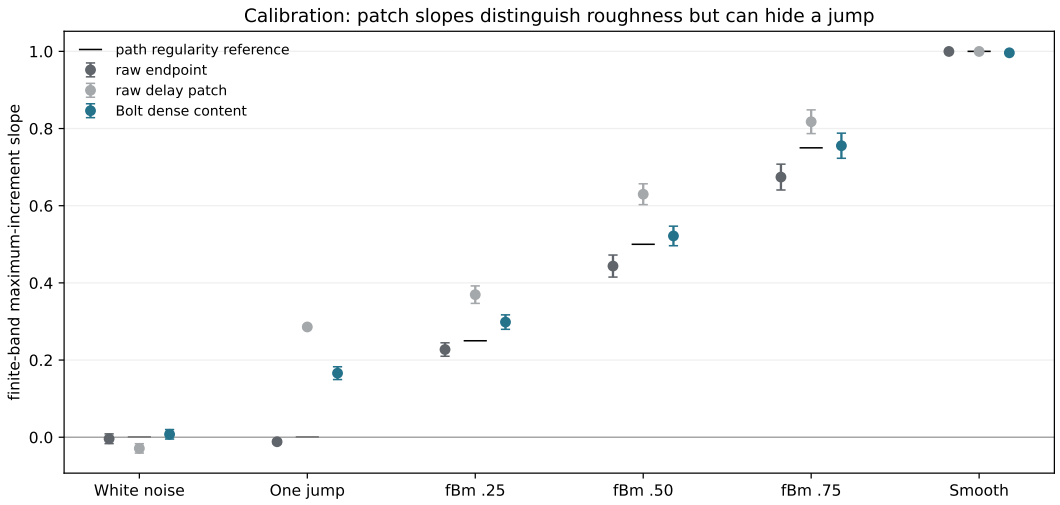

Figure 2: **Calibration of finite-band slopes.** Twelve realizations per scenario, with mean ± sample SD. The ordering of fractional-Brownian roughness is visible, but the positive slope of the jump control after patch encoding prevents interpreting the regression as an asymptotic certificate.

In Figure 2, the raw fBm slopes are $0.227, 0.444, 0.674$, and dense Bolt gives $0.298, 0.522, 0.755$; smooth and white-noise Bolt slopes are 0.996 and 0.008. These values support sensitivity to roughness while exposing finite-scale and representation bias. Crucially, the jump gives 0.166 ± 0.017 after dense Bolt, compared with -0.012 ± 0.005 before encoding: distributing a transition across patch coordinates changes its observed scaling. On the matched band of raw lags at least 16, this same dense jump representation gives -0.013 ± 0.006, exposing the scale crossover. Positive fitted slopes therefore need refinement checks, not a universal bias correction.

For a complementary check, restrict each fixed dense feature path, without interpolation or re-encoding, to nested grids of $n = 65, 129, 257, 513$ points. On each grid compute the exact all-pair seminorm

$$A_n(\alpha) := \max_{i<j} \frac{d^{-1/2}\|z_j - z_i\|_2}{|t_j - t_i|^\alpha}, \qquad \alpha \in \{0.1, 0.25, 0.5\}. \tag{241}$$

The grid endpoints span the largest power-of-two number of sample intervals inside the valid dense path, centered in that path. For the retained lengths this is one half of the raw window; times remain normalized by $N_{\mathrm{raw}} - 1$. Every positive lag and every starting position on each grid enter the maximum. The grids are nested, so A_n cannot decrease. The 513-point cap reaches raw spacing one for Beijing and traffic, four for appliances and the two ETT channels, and 32 for Jena. This check therefore does not reach the finest measurement cadence in every domain, nor cover the whole observation interval.

In Figure 3, at $\alpha = 0.25$, dense traffic and Beijing have mean A_{513}/A_{65} of 1.045 and 1.129. Dense weather and appliances still grow by factors 1.626 and 1.674. Thus some continuous representations have encouraging prefactor stability over these grids, but the observations do not establish a shared 0.25-Hölder class. We extract no pass/fail exponent.

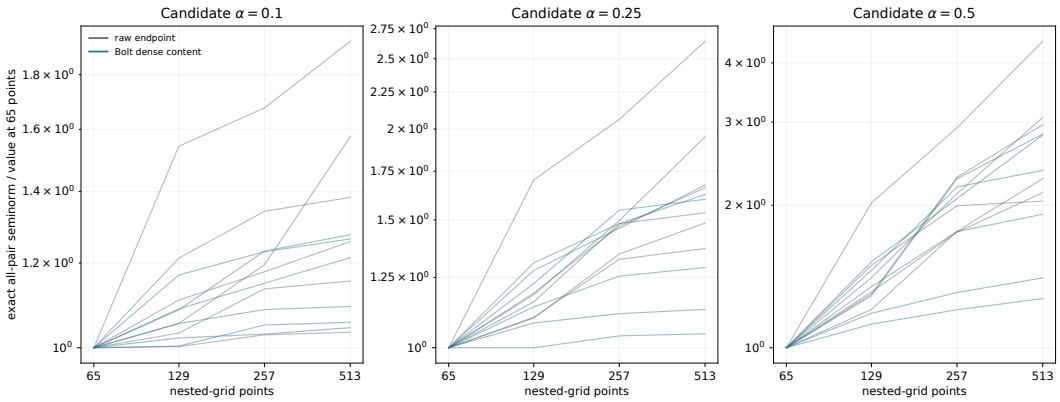

Figure 3: **Exact candidate Hölder prefactors on nested restrictions.** Each thin curve is a domain's mean ratio $A_n(\alpha)/A_{65}(\alpha)$, for raw endpoints or dense Bolt content. Domains are not pooled as independent samples. The finite cadence and central subinterval limit the scope of this stability check.

H.7 A STRUCTURED-TEXT CORPUS SCREEN

We extend the original four-source input baseline with three candidates chosen before measuring their increment curves: Shakespeare drama and verse, the King James Bible as formulaic historical prose, and a Python-code control. The motivation is to test whether more structured textual sources exhibit stronger geometric regularity. All three outcomes are retained; neither sources, windows, nor fitting bands are selected for favorable slopes.

Data and unchanged embedding. The public texts are the *Tiny Shakespeare* corpus and Project Gutenberg's King James Bible.[2] We read at most one million bytes of each source, remove the Gutenberg header, and retain the first four disjoint $N = 16{,}384$-token windows. The Python source is a sorted snapshot of Git-tracked files under `code/`, including vendored files, concatenated once. The vendored files were included because top-level code alone was too short; this adjustment preceded measurement of any candidate's increments. Neither the new public texts nor the Python stream is periodically repeated. Only token windows, source hashes, and small statistics are stored, not full corpora or embeddings.

Every vector uses the same frozen BigBird $d = 768$ input map as the original baseline: word and token-type embeddings plus the learned position table linearly interpolated to N positions, followed by the pretrained LayerNorm. No attention, training, or smoothing is applied. A paired control permutes each new window's content IDs before this map, keeping the same positional grid. Fits use the same nine dyadic lags $r = 1, \dots, 256$ satisfying $r/(N-1) \le 1/32$. In addition to the maximum, we fit the RMS increment $M_2(r) := [(N-r)^{-1} \sum_k d^{-1} \|z_{k+r} - z_k\|_2^2]^{1/2}$. Table 5 reports all seven sources; SD describes four windows per source, not independent documents or corpus-level uncertainty.

Outcome and scope. The new maximum-slope means are 0.0106, 0.0115, and 0.0124: none exceeds the original IMDb mean 0.0130. Shuffling produces similarly small values, and mean RMS slopes remain between 0.0148 and 0.0191 across all seven sources. Thus textual structure or formulaic syntax does not by itself make these subword embeddings geometrically regular. The first-four-dyad and expanded $h \le 1/8$ fits are also retained; their largest admissible dyads are 8 and 1024, respectively, since $2048/(16384-1) > 1/8$. No alternative band supplies a strong positive regime in these additional sources. This remains a finite-scale negative result for one input pipeline, not a statement about every possible text representation. The separate layer-wise study below investigates whether processing changes this geometry.

[2] Sources: `char-rnn/data/tinyshakespeare` and Project Gutenberg, ebook 10. Download-prefix hashes and exact retained token IDs are saved with the results.

Table 5: **The text screen does not reveal a more regular input regime.** Mean ± sample SD on the fixed fine band, using identical full-input BigBird embeddings. Shuffled controls are measured for all three new sources; dashes denote unmeasured controls at this long context. The original LaTeX stream is a repeated technical control, as described above.

Source	$\widehat{\alpha}_\infty$	$\widehat{\alpha}_2$	Shuffled $\widehat{\alpha}_\infty$
WikiText-103	0.0122 ± 0.0012	0.0188 ± 0.0006	–
AG News	0.0126 ± 0.0018	0.0191 ± 0.0003	–
IMDb	0.0130 ± 0.0008	0.0180 ± 0.0001	–
LaTeX control	0.0119 ± 0.0012	0.0148 ± 0.0004	–
Shakespeare drama	0.0106 ± 0.0018	0.0173 ± 0.0002	0.0109 ± 0.0017
KJV Bible	0.0115 ± 0.0005	0.0178 ± 0.0001	0.0117 ± 0.0024
Python control	0.0124 ± 0.0007	0.0150 ± 0.0003	0.0103 ± 0.0027

The experiment is reproduced by `code/text_corpus_holder_screen.py`: `--prepare` freezes small token inputs, `--run` evaluates the cached input map, and `--report` reanalyzes saved increments. The executed `code/holder_increment_comparison.ipynb` regenerates the publication panels and the full screen report without new inference.

H.8 Text representations across transformer depth

Nearly flat input-embedding curves need not describe the geometry of internal representations. We therefore repeat the same multiscale diagnostic after every transformer block, asking whether depth improves the effective Hölder scaling of a fixed text window. The main observation is a modest but consistent intermediate-layer gain in BigBird, rather than a monotone increase through the network.

Matched windows and native inference. We evaluate the frozen 12-block BigBird encoder (Zaheer et al., 2020) at its native 4096-token context and the six-block causal DistilGPT2 decoder from the GPT-2 family (Radford et al., 2019) at 1024 tokens.[3] For each model and each of WikiText-103, AG News, IMDb, and a frozen LaTeX source-tree control, we retain the first eight nonoverlapping full content-token windows. The technical control includes retained source variants and is concatenated once, without periodic repetition. Public documents are joined with the tokenizer's separator. A paired, seeded control permutes content IDs before inference, including internal separators but preserving BigBird's outer markers. The two models use different tokenizers and window lengths; only within-model comparisons are paired.

All depths use exactly the same windows and lags, with no training, smoothing, or positional interpolation. BigBird's two added outer markers are excluded from measurement, leaving $N = 4094$ content vectors; DistilGPT2 has $N = 1024$. Both have increment dimension $d = 768$. Depth zero denotes the embedding representation. We record every block output and, for DistilGPT2, its final LayerNorm separately from block six. Primary fits use (24), with $h_r := r/(N-1)$: seven dyadic lags through $r = 64$ for BigBird and five through $r = 16$ for DistilGPT2. These native-length results are distinct from the earlier 16,384-token input-only baseline.

Comparing multiscale shape across depth. Let $M^{(w)}_{\infty,\ell}$ denote (23) in window w at depth ℓ. Figure 4 overlays the normalized mean curves

$$Q_\ell(r) := \frac{\overline{M}_{\infty,\ell}(r)}{\overline{M}_{\infty,\ell}(1)}, \qquad \overline{M}_{\infty,\ell}(r) := \frac{1}{8}\sum_{w=1}^{8} M^{(w)}_{\infty,\ell}(r). \tag{242}$$

Dividing by a positive lag-independent constant does not change a log–log slope. This normalization separates changes in curve shape from changes in amplitude; the tables still summarize fits made separately in each window, not fits to the displayed mean curves. The magnified vertical range is important: visibly different curves here often differ by only 10–20% across hundreds of token lags.

[3] Pinned snapshots: BigBird `5a145f7` and DistilGPT2 `2290a62`. Inference uses `transformers==4.57.6`, CPU float32, and evaluation mode. BigBird's unmodified block-sparse evaluation helper returns block-zero indices rather than sampled random links; these are links to block zero, not zero attention weights. We impose no custom attention pattern.

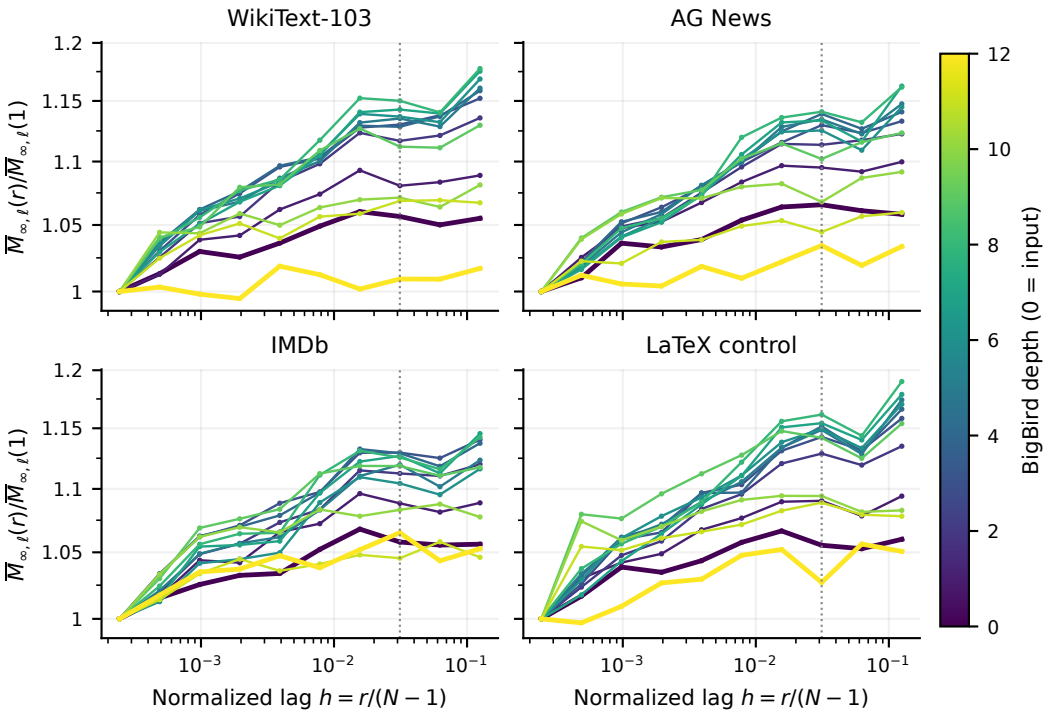

Figure 4: **BigBird's multiscale geometry varies with depth.** All depths are overlaid on the same log–log axes, using the same eight windows per source. Curves are ratios of mean worst increments as in (242); purple and yellow mark the input and final block. The dotted line is the fixed fine-band cutoff $h = 1/32$. Intermediate blocks have steeper finite-band curves, but the gain does not persist to the final block.

An intermediate-layer gain in BigBird. The mean primary slope rises from 0.013–0.014 at the input to 0.028–0.033 around blocks seven and eight (Table 6). Block eight improves on the input in all eight retained windows of each source on the primary band. The mean gain remains positive on the first-four-dyad and expanded $h \leq 1/8$ bands, although not every short-band paired difference is positive. Thus the improvement is not merely a visual normalization effect. It is non-monotone: final-block means return to 0.002–0.014. The mean depth profiles peak at block eight for the three natural-language sources and block seven for LaTeX; these peak locations are descriptive, not choices used to tune the fitting band.

Table 6: **Effective maximum-increment slopes across depth.** Mean ± sample SD across eight matched natural-order windows per source, on $h \leq 1/32$. BigBird's final state is block 12; DistilGPT2's is the output after the separate final LayerNorm. SD is descriptive, not a confidence interval: windows may share documents or repeated content.

BigBird (12 bidirectional blocks)			
Source	Input	Block 8	Final
WikiText-103	$.0130 \pm .0016$	$.0308 \pm .0027$	$.0023 \pm .0060$
AG News	$.0141 \pm .0015$	$.0306 \pm .0042$	$.0039 \pm .0082$
IMDb	$.0143 \pm .0014$	$.0279 \pm .0037$	$.0106 \pm .0082$
LaTeX control	$.0143 \pm .0019$	$.0320 \pm .0043$	$.0140 \pm .0086$
DistilGPT2 (6 causal blocks)			
Source	Input	Block 6	Final
WikiText-103	$.0165 \pm .0105$	$.0108 \pm .0399$	$.0161 \pm .0152$
AG News	$.0176 \pm .0072$	$.0171 \pm .0328$	$.0129 \pm .0204$
IMDb	$.0174 \pm .0093$	$-.0105 \pm .0293$	$-.0002 \pm .0187$
LaTeX control	$.0227 \pm .0096$	$.0150 \pm .0361$	$.0206 \pm .0256$

The gain is not specific to intact linguistic order. Shuffled-input BigBird slopes at block eight are $0.0363, 0.0382, 0.0354, 0.0291$, respectively (Figure 5). This does not isolate the effects of learned weights, positions, and sparse-attention locality, but it prevents attributing the intermediate gain solely to meaningful word order. The final layer also contracts amplitudes: its centered token RMS is 0.55–0.70 times the input value, whereas the adjacent maximum is 0.82–0.90 times its input value. Smaller increments therefore need not imply a larger scaling exponent.

Causal representations and average-increment scaling. DistilGPT2 does not show a positive mean final-minus-input change in the primary maximum slope for any source. For comparison, we fit the same slopes to the root-mean-square increment

$$M_{2,\ell}(r) := \left(\frac{1}{d(N-r)} \sum_{k=0}^{N-r-1} \| z_{k+r}^{(\ell)} - z_k^{(\ell)} \|_2^2 \right)^{1/2}.$$

This statistic exhibits a stronger depth effect: input means are 0.0016–0.0062, while final means are $0.0216, 0.0393, 0.0569, 0.0661$. The corresponding shuffled means are $0.0231, 0.0218, 0.0293, 0.0347$. Thus natural order improves this average statistic except on WikiText, but an RMS gain does not establish the uniform increment control used in the theory. Figure 5 keeps the two statistics distinct and shows the final normalization separately; it is not a seventh transformer block.

Sensitivity and scope. Removing the first and last 64 content positions after inference, without changing the context or grid denominator, leaves final maximum-slope means between 0.002 and 0.020 for BigBird and between -0.001 and 0.017 for DistilGPT2. Alternative bands and percentile/RMS statistics are retained in the saved results. Exact-lag maxima need not be monotone; negative fits to nearly flat curves are not negative Hölder exponents. A post-run cumulative maximum over tested dyads is also available, but it is not the supremum over all integer lags. These fixed-context experiments support depth-dependent *effective* regularity, including a genuine intermediate gain, not asymptotic Hölder improvement under refinement or a common all-resolution bound. No prediction-quality claim is made.

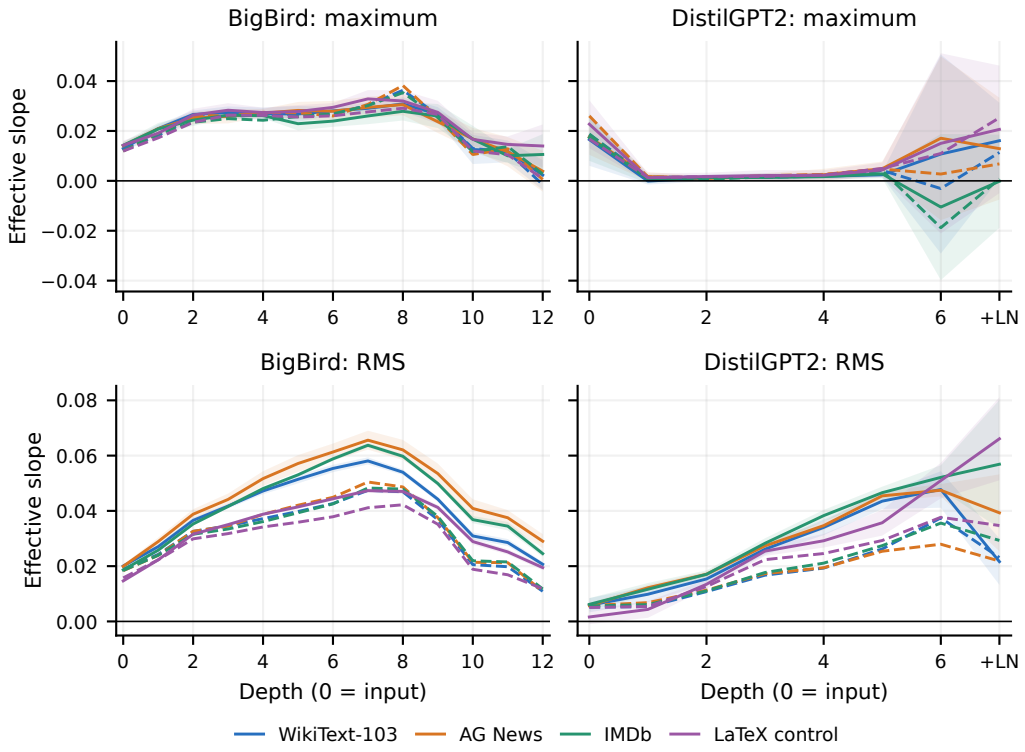

Figure 5: **Depth, token order, and the choice of increment statistic.** Mean fine-band slopes at every block for BigBird and causal DistilGPT2. Solid lines use natural order, dashed lines use shuffled inputs; shading is one sample SD for natural-order windows. Top: maximum increments. Bottom: RMS increments, a weaker statistic. DistilGPT2's +LN point applies the final normalization to block six. Architectures, tokenizers, and context lengths differ, so cross-model differences do not isolate an effect of masking.